生物多样性译丛

（三）

中国科学院生物多样性委员会

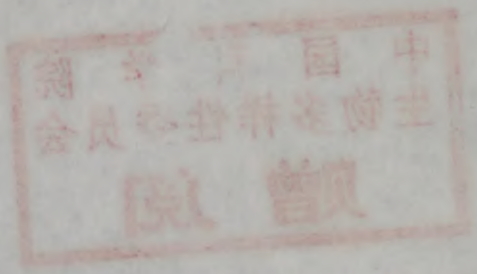
科学出版社

生物多样性译丛（三）

中国科学院生物多样性委员会

科学出版社
1997

(京)新登字 092 号



生物多样性译丛(三)
中国科学院生物多样性委员会
科学出版社出版
北京东黄城根北街 16 号
邮政编码:100717
兵器工业出版社印刷厂印刷
新华书店北京发行所发行　各地新华书店经售
*
1997 年 12 月第一版　开本:787×1092 1/16
1997 年 12 月第一次印刷　印张:14.25
印数:1—1000 字数:410 千字
ISBN 7-03-005669-8/Q・616
定价:41.00

译　序

随着时间的推移，生物多样性的重要性越来越为社会各界所理解和重视。众所周知，生物多样性中最为重要的有物种多样性、生态系统多样性和遗传多样性三个层次。遗传多样性是这三个层次中研究相对较为薄弱的，在我国尤其如此。

遗传多样性研究具有十分重要的意义。其研究结果可为物种濒危机制的探讨、种质资源的保存与利用、生物分类与演化的研究、生态适应机制的探讨等提供重要证据或依据。我国的遗传多样性研究取得了重要进展，特别关于大熊猫濒危机制的研究受到国外同行的关注。但是，除中国科学院生物多样性委员会“生物多样性研究丛书”之一《中国动植物遗传多样性》(胡志昂和张亚平主编)以外，国内还没有遗传多样性方面的专著出版。有鉴于此，我们组织翻译了“分子进化基础”(中科院昆明动物研究所细胞与分子进化开放研究实验室陈建华译，张亚平、吴春花和李海鹏校，其中序和前言由赵丽惠译，马克平校)和“分子变异与生态学问题(中科院植物研究所魏伟译，钱迎倩校)以补充国内这方面资料的不足。为了促进生物多样性研究的重要支撑学科之一，生物系统学的发展，我们将“生物系统学议程 2000 年”(中科院动物研究所郭寅峰译，中科院植物研究所钱迎倩和动物所周红章校)同时编入此书，作为《生物多样性译丛》之三正式出版。其它三集分别是：《生物多样性译丛一》(包括《保护世界的生物多样性》和《生物多样性一有关的科学问题与合作研究建议》两部分)、《全球生物多样性策略》(作为《译丛》之二)和《生物多样性公约指南》(作为《译丛》之四)。我们将继续选择国外本领域好的资料翻译出版，及时把国外新的理论、方法和动态介绍到国内，以推动中国的生物多样性保护和研究工作。

北京教育学院刘玉明先生和中科院植物所赵丽惠同志认真校订译稿，北京海岸文化服务中心刘万海先生大力支持。在此，向他们表示衷心的感谢。

马克平

1997 年 7 月 25 日

目　录

《分子进化基础》

《分子变异与生态学问题》

《2000 年系统学议程:制订生物圈计划》

分 子 进 化 基 础

原作：李文雄（美国得克萨斯大学）
D. 戈劳尔（以色列特拉维夫大学）
翻译：陈建华

Fundamentals of Molecular Evolution

Wen-Hsiung Li (The University of Texas, Houston)
Dan Graur (Tel Aviv University)

Sinauer Associates, Inc., Publishers
Sunderland, Massachusetts, 1991

序

1869 年，Johann Friedrich Miescher 发现了 DNA。他曾经“大胆”地认为，DNA 可能和遗传有关。但过了一段时间，他还是放弃了这种“荒谬”的观点，并花费 30 多年的时间去研究鱼精蛋白—精子细胞中非常基本的蛋白质组分。按照 Miescher 的观点，这种蛋白质是遗传的关键所在。后来发现，DNA 分子不仅仅是遗传的关键，正如一位分子进化学先驱 Emile Zuckerkandl 所说的那样，它们还是“进化史的文献资料”。实际上，每一种活组织的 DNA 都是历史过程的积累。然而，包容在这些过程里的信息处在一种杂乱无章的无序状态中，是分散的或片段的。有一些信息是隐藏的或是“伪装”的，很难辩认出来；有一些已丢失，毫无踪迹可寻。分子进化的目的就是要澄清这些历史过程，整理信息使之有序，读出并解译信息。

由于进化过程难以用基因材料研究清楚，用分子资料不仅可以重建进化年代，而且也可以弄清进化过程的驱动力。分子生物学的重大突破，如基因克隆技术、DNA 序列分析、限制性核酸内切酶片段分析等，在某种意义上，已使科学家们处于一种新的十分有利的位置。我们能够洞察一个从未见过的世界，在这里，基因是通过复制、DNA 混匀、核苷酸替代、转移及基因转化进化的。在这个世界里，基因组或静止，或流动，有时在过了很长时间后会有微小的变化，有时又会在眨眼间发生戏剧性的地质学尺度的变化。

通过研究遗传材料，我们也能试着将生物界按系统发育事实来划分，并建立分类系统。与传统的研究进化方法相反，例如比较解剖学、形态学和古生物学，都毫无必要地将自己限制在极端相似组织的进化关系研究中。我们现在已经能够建立巨大的家系树，它连接脊椎动物、昆虫、植物、真菌和细菌，并能够追溯它们的共同祖先，一直到无法追溯的时代。

本书的宗旨是从分子水平上描述进化的原动力，进化过程的驱动力及其基因组，基因和它们的产物长期进化的各种分子机制所产生的结果。另外，本书从进化观点为分子材料的比较和系统发育分析提供了基本的方法论。尽管个体受自然选择和其它过程所影响，我们依然强调：是群体和基因随着进化时间而变。为了在这些看起来完全不同的水平间构造联系，我们借用了群体遗传学的基本概念。

我们着手为分子进化“初学者”写了这本书。同时，也努力保持了方法的科学和规范，并包含了一些定量方法。因此，可以用数学和直觉解释，从分子水平上描述进化现象和机制。这两种解释都不会削弱彼此的说服力，还会相互补充，以帮助读者更好地领会观点。我们没有试图达到泛泛的完满性，而是提供了大量实例，以支持和澄清许多理论观点和有争议的问题。

许多年来，生物化学家和分子生物学家们把进化研究看作是大胆推测，无根据的假设和无学科方法论的统一体，而这种评价从未正确过。分子方法的引进不容置疑地将进化论变成一种“硬”科学。在这里，能够从经验数据测度、计数或用计算机处理相关的参数，也能用事实去检验理论的正确性。在今天的进化研究中，推测学也服务于同物理学同样的目的。它们是定量性假说，以刺激实验性工作。由此，理论能够得以校正、精炼或被驳斥。本书的主要目的之一是通过加强本领域的事实基础，表明进化研究已经达到了 William Herschel 先生于 1831 年所说的所有自然科学的共同目的，即可以这样陈述其主张：“不是模糊和泛泛的，而是在位置、重量和量度上尽可能精确”。

我们感谢许多同事、学生和朋友，是他们帮助我们编辑了本书。他们的评论、建议、指正和讨论大大地提高了本书的质量，也避免了许多。我们荣幸地感谢 Sara Barton，Adina Breiman，David Cutler，David Hewett-Emmett，Winston Hide，Austin Hughes，Li Jin，Margaret Kidwell，Amanda Ko，Giddy Landan，William S. Lewis，Volker Loeschcke，Ora Manheim，David Mindell，Tatsuya ota，Lori Sadler，Paul Sharp，Yuval Shuali，Jurgen Tomiuk，Danid Wool 和 Chung-I Wu. 特别地尴尬，我们感谢 C. William Birky Jr. 和 Bruce walsh，他们两位从头到尾地校阅了文稿并提出了很多有益的建

议。我们也向 Masatoshi Nei 博士表达感激之情，他对我们的调查研究工作给予了极大的鼓励和指导。来自世界卫生组织和美国—以色列两国科学基金会的支持也是本项合作成果得以面世的原因之一。

李文雄

D·戈劳尔

前　言

什么是分子进化

分子进化包括两个研究领域：(1) 大分子进化；(2) 基因和有机体进化史的重建。关于“大分子进化”，我们指的是在进化年代中，出现在遗传材料（如 DNA 序列）中的变化速率、模式及产物（如蛋白质）以及与此变化相关的机制，也被称做分子种系发生史。旨在探讨有机体和大分子进化史（正如从分子材料所推知的那样）。

这两个研究领域似乎由不相关的研究项目组成，因为第一个研究领域的目的是要阐明分子进化的原因和结果；而第二个研究领域是仅仅把分子用作工具，以重建有机体及其遗传组分的演化史。然而在实践中，这两个方面是紧密相关的，一个领域的进步也会促进另一领域的发展，例如：种系发生知识对决定分子特征的变化顺序是很必要的；相反，给定分子的变化模式和速率知识对试图重建类群的进化史是很重要的。

传统上，第三个研究领域，前生物（Prebiotic）进化或称“生命起源”也包含在分子进化框架中。然而，它涉及到很多猜想，也极少用到定量方法，并且，在前生物系统（即缺乏可复制基因的系统）信息传递过程的规律，目前还不清楚。因此，本书并不侧重生命起源，有兴趣的读者可以参阅 Oparin (1957)，Cairns-Smith (1982)，Dyson (1985) 和 Loomis (1988) 的著作。

分子进化研究起源于两个不同的学科：群体遗传学和分子生物学。群体遗传学为研究进化过程提供了理论基础，而分子生物学则为其提供了经验数据。因此，掌握群体遗传学和分子生物学的基本知识对理解分子进化是很有裨益的。

1 基因结构与突变

本章提供为研究DNA水平上的进化过程所必需的、关于分子生物学方面的基础知识。最为基本的部分是一种典型的真核生物基因的基因组结构和各种突变类型。更进一步的背景知识将在以后的有关章节中讲述出。某些术语请参看词汇表。

1.1 DNA序列

除某些病毒外，所有生物的遗传信息都由脱氧核糖核酸DNA(deoxyribonucleic acid)分子所携带。DNA通常是由两条相互缠绕的互补链所组成，形成一种右手螺旋状。每条链都是一条由四种核苷酸组成的线性多聚核苷酸。有两种嘌呤(purines)：腺嘌呤A(adenine)和鸟嘌呤G(guanine)，两种嘧啶(pyrimidines)：胸腺嘧啶T(thymine)和胞嘧啶C(cytosine)。两条链靠核苷酸之间的氢键联系在一起。腺嘌呤与胸腺嘧啶通过两个氢键，又称弱键(weak bond)配对；而鸟嘌呤与胞嘧啶通过三个氢键，又称强键(strong bond)配对(图1—1)。

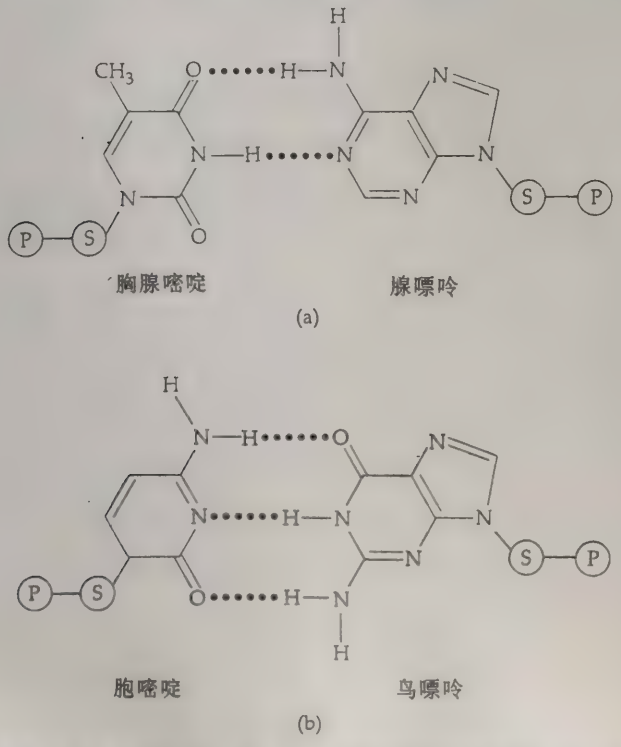


图1—1 互补碱基靠氢键(由圆点构成的线)配对，(a)在胸腺嘧啶和腺嘌呤之间(弱键)，和(b)在胞嘧啶和鸟嘌呤之间(强键)。P：磷酸根；S：糖。

DNA序列中的每一个核苷酸都包含一个戊糖(脱氧核糖)、一个磷酸基团和一个嘌呤或嘧啶碱基。DNA分子的主干由糖和磷酸等部分构成，它们通过不对称的5′—3′磷酸二酯键共价连结在一起。所以，DNA分子是一种有极性的分子，它的一端，在末端核苷酸的5′碳原子上有一个磷酸根(—P)，另一端，其末端核苷酸的3′碳原子上有一个游离羟基(—OH)。磷酸二酯键的方向决定着分子的特性；

例如，序列 5′－G－C－A－A－T－3′与序列 3′－G－C－A－A－T－5′ 是两种不同的序列。为了方便起见，DNA 序列按其转录的顺序，即从 5′端到 3′端书写，它们又分别被称为上游(upstream)和下游(downstream)方向。DNA 的双螺旋形式有两条链，按反向平行方向排列着(图 1－2)。

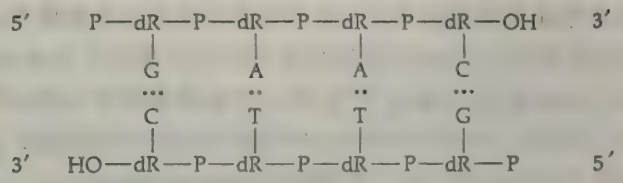


图 1－2　双链 DNA 的反向平行结构的图示

P，磷酸根；dR，脱氧核糖；CH，羟基；A，腺嘌呤；G，鸟嘌呤；C，胞嘧啶；T，胸腺嘧啶；—，共价键；· ·，弱键；⋯，强键。

核糖核酸 RNA(ribonucleic acid)既有双链分子又有单链分子。RNA 与 DNA 不同之处在于，其主干糖的部分是核糖而不是脱氧核糖，由尿嘧啶核苷酸代替了胸腺嘧啶核苷酸。腺嘌呤，胞嘧啶，鸟嘌呤，以及胸腺嘧啶/尿嘧啶，都是标准核苷酸。某些功能 RNA 分子，特别是 tRNA 分子，常含有非标准核苷酸，即在转录之后导入 RNA 分子中的、将标准核苷酸加以化学修饰的核苷酸。

1.2　基因结构

过去，基因被定义成为一个多肽链编码或确定一个功能 RNA 分子的一个 DNA 片段。然而，最近的分子研究却从根本上改变了我们对基因的认识，所以，我们将采用一个更有点生气的定义，即基因是对某一特殊功能至关重要的基因组 DNA 或 RNA 的一个序列。执行这一功能可能并不需要该基因被翻译出来，甚至无需被转录出来。

现在，已有三类基因被认识：(1)为蛋白质编码的基因(protein-coding genes)，它们在被转录成 RNA 之后再被翻译成蛋白质；(2)RNA 基因(RNA-specifying genes)，它们仅被转录；(3)调节基因(regulatory genes)。按照狭义的定义，第三类只包括那些不转录的序列。被转录的调节基因本质上属于前两类之一。蛋白质编码基因和 RNA 基因又被看成是结构基因(structural genes)。注意，某些作者将结构基因的定义，限制在只包括蛋白质编码基因的范围。

在细菌中，结构基因的转录仅由一种依赖 DNA 的 RNA 多聚酶执行。与之不同的是，在真核生物里要用到三种 RNA 多聚酶。核糖体 RNA(rRNA)基因由 RNA 多聚酶Ⅰ转录，蛋白质编码基因由 RNA 多聚酶Ⅱ转录，小分子细胞质 RNA(scRNA)基因，象转移 RNA(tRNA)的基因，则由 RNA 多聚酶Ⅲ转录。某些小分子细胞核 RNA(snRNA)基因由多聚酶Ⅱ转录，另一些则由多聚酶Ⅲ转录。一种 snRNA 基因 U_6，可能既被多聚酶Ⅱ又被多聚酶Ⅲ转录。

蛋白质编码基因

真核生物的一个标准的蛋白质编码基因由转录部分和不转录部分构成(图 1－3a)。不转录部分根据它们相对于蛋白质编码基因的位置，被命名为 5′和 3′侧序列(flanking sequences)。5′侧序列含有几种信号(特异序列)，它们决定转录过程的起始、缓急和终止。因为这些调节序列起始转录过程，所以，它们也被称为启动子(promoters)，而它们所处的部位则叫启动子区(promoter region)。启动子区由以下信号构成：TATA 块(TATA box)，位于转录起始点上游 19－27 碱基对(bp)处，更上游一些的 CAAT 块(CAAT box)，以及一个或多个 GC 块(GC box)拷贝，后者由序列 GGGCGG 或其变异体组成，位于 CAAT 块的周围(图 3a)。可能在定向方面起作用的 CAAT 块和 GC 块，控制 RNA 多聚酶的起始联结，而 TATA 块则控制转录起始点的选取。不过，我们注意到，以上信号对启动子的功能而言都不是必不可少的。某些基因不具有 TATA 块，所以没有一种独一无二的转录起始点。另一些基因既没有 CAAT 块也没有 GC 块，它们的转录起始是受 5′侧区域中的其他因素所控制的。3′侧序列含有关于转录过程终止的信号和多聚腺苷酸尾部的信号。目前，还不能精确地标明一个基因开始和终止的位

点。

真核生物中为蛋白质编码的基因的转录起始于转录起始部位(transcription-initiation site)(RNA转录物中的帽子部位(cap site)),结束于终止部位(termination site),后者与成熟信使RNA(mRNA)分子的多聚腺苷化作用(polyadenylation)或多聚腺苷酸尾部位(poly(A)-addition site)可能等同,也可能不等同。换言之,转录的终止可能发生在比多聚腺苷酸尾部更下游的地方。转录而成的RNA,又称前信使RNA(pre-messenger RNA)(pre-mRNA)包含5′和3′不翻译区,(untranslated regions),外显子(exons)和内含子(introns)。内含子,或插入序列,是一些虽被转录但在前mRNA分子的加工中被剪除的序列。所有保留在经拼接而达成熟的mRNA中的基因组序列称外显子。被翻译出来的外显子或外显子的某些部分称为蛋白质编码的外显子或编码区(coding regions)。

有几种类型的内含子,按它们从前mRNA中剪除的特异机制来界定其性质(见Lewin,1990)。这里我们只关心那些受RNA多聚酶Ⅱ转录的细胞核基因中的内含子。这些内含子在分子成熟期间,在酶促作用下从前mRNA中剪除。这些内含子的拼接位点(splicing sites or junctions)可能由每一内含子的5′和3′端上的核苷酸来决定,分别称给体(donor)和受体(acceptor)部位。例如,所有真核生物的细胞核内含子都以GT开始,以AG结束(GT－AG规则(GT-AG rule)),而已表明这些序列对于正确的剪切和拼接是至关重要的(图1－3a)。邻近内含子的外显子序列可能对拼接部位的决定也有贡献。此外,每一内含子都含有一段特别的序列,TACTAAC框(TACTAAC box),位于内含子3′端上游

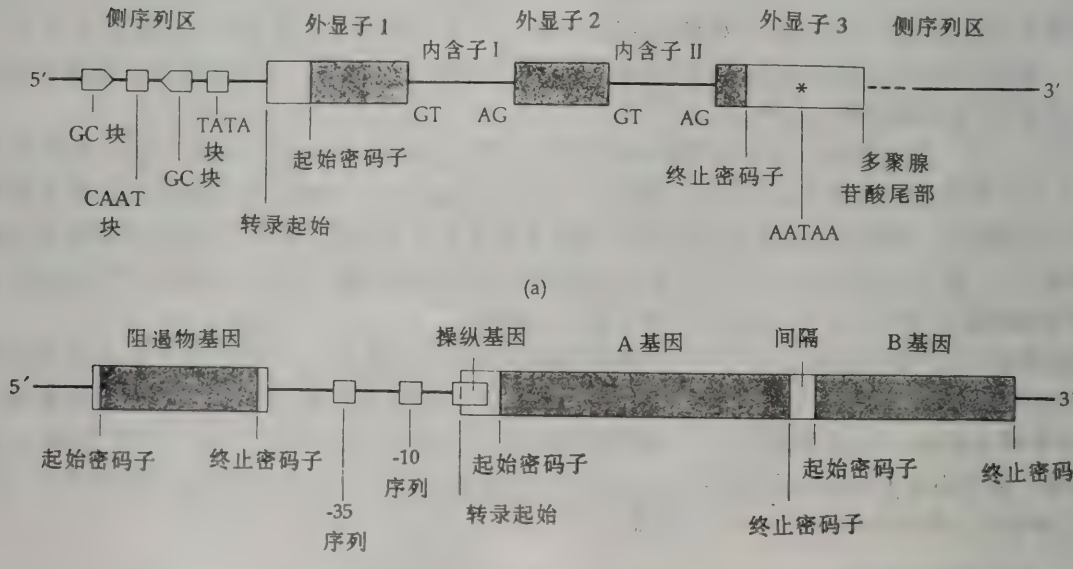


图1－3 (a)典型真核生物的蛋白质编码基因的图示结构。注意,按照习惯5′端放在左边。(b)受诱导的原核生物的操纵子的图示结构。基因A和B是为蛋白质编码的基因,并被转录形成一个信使RNA。阻遏物基因为一个阻遏蛋白编码,后者结合在操纵基因上,并通过遏制RNA多聚酶的移动从而阻碍结构基因的转录。操纵基因是一段至少有10个碱基长的DNA区域,它可能与操纵子中基因的被转录区域重叠。经与一种诱导物(一种小分子)结合后,阻遏物转化成一种不能与操纵基因结合的形式。于是RNA多聚酶就能起动操纵子中基因A和B的转录(见Lewin 1990)。(a)和(b)两图中的各区都不是按其尺度来画的。

约30个碱基的地方。这一序列在酵母细胞核基因中是相当保守的,而它在更高等的真核生物的基因中则要变化得多样一些。拼接过程包括5′拼接部位的裂开,以及内含子5′端的G和TACTAAC框中第6位上的A之间形成磷酸二酯键。随后,3′拼接部位裂开,而两个外显子则被连在一起。

内含子的数目因基因而异。某些基因具有十几个内含子,其中有些内含子可能有几千个核苷酸长。另一些(例如,大多数组蛋白基因)则完全没有内含子。外显子在整个基因中并不是均匀分布的。

有些外显子密集在一起，另外一些则位于与邻近外显子相距很远的地方。图 1－4 给出了说明这一现象的一个例子。注意，细胞核蛋白质编码基因的绝大多数具有内含子。并非所有内含子都起分裂编码区的作用。有些内含子出现在不翻译区域，主要在转录-起始位点和起始密码子之间的区域中。

真细菌类的蛋白质编码基因与真核生物中的这类基因有几方面不同，最重要的是，它们不含内含子（图 1－3b）。真细菌中的启动子含有一个－10 序列（－10 sequence）和一个－35 序列（－35 sequence），分别位于转录起始位点上游 10bp 和 35bp 的地方. 前者也称普里伯劳块（Pribnow box）具有序列 TATAAT 或其变异形式，而后者则具有序列 TTGACA 或其变异形式。原核生物的启动子在离－35 序列更上游的地方，可能还含有其他一些特别的序列。

原核生物中几个结构基因可能呈连续排列以形成一个基因表达单位，它被转录成一个 mRNA 分子，接着再翻译成不同的蛋白质。这种单位通常都含有控制属于该单位的基因协调地表达的遗传因素。这一由基因排列成的整体称为一个操纵子（operon）（图 1－3b）。

RNA 基因

原核生物和真核生物中的 RNA 基因，其结构通常是相似的。这些基因一般不含内含子。然而，在有些生物中，象纤毛虫、粘菌和细菌中，确定 RNA 的基因就可能含有在 RNA 变成有功能分子前必须被剪除的内含子。涉及某些 RNA 基因的转录调节的序列因素，有时被包括在决定功能终产物的序列的范畴内。特别地，所有确定 tRNA 的基因都含有一个受 RNA 多聚酶Ⅲ识别的内部转录起始点。

许多 RNA 分子在转录之后都受到修饰。这类修饰包括标准和非标准核苷酸的掺入，标准核苷酸转变成非标准核苷酸的修饰，向 5′或 3′末端的核糖核苷酸的终端序列的酶促添加等。

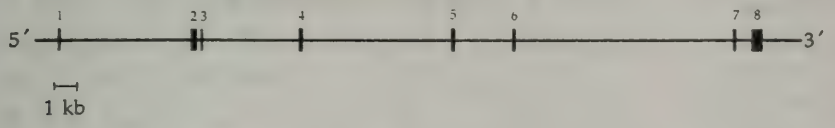


图 1－4　人的Ⅸ因子基因中 8 个外显子的位置。图中竖线代表 8 个外显子。本图只标出了被转录的区域，外显子和内含子均按其尺度绘出。注意，5′不翻译区比 3′不翻译区短。资料取自 Yoshitake 等（1985）。

调节基因

我们对调节基因的认识比对其他类型基因的认识要晚一些。暂时只有几种这类基因或基因家族被鉴别出来。它们是：（1）复制子基因（replicator genes），作为 DNA 复制的起始和终止的特异位点而起作用，（2）重组子基因（recombinator genes），提供受与重组有关的酶识别的特异位点，（3）分离子基因（segregator genes），在有丝分裂和减数分裂期间，提供将染色体附着到分离机构上的特异位点，以及（4）附着位点（attachment sites），为蛋白质、激素，或其他分子所附着。可能还存在更多的调节基因，它们中有些可能在功能和位置两方面都是独立于结构基因的，并且可能与更复杂的调节功能，象多细胞生物的个体发育有关。

1.3　遗传密码

蛋白质合成涉及到一个解码过程，藉此，一个由 mRNA 分子携带的遗传信息通过转移 RNA（tRNA）介体的使用，而被翻译成氨基酸。20 种基本氨基酸及其缩写形式列于表 1－1。翻译从翻译起始位点开始，进行到一个终止信号时为止。翻译过程涉及相互邻近的不重叠核苷酸三联体，称为密码子（codons）的连续识别。一个将被翻译的序列的相位由起始密码子决定，称为阅读框架（reading frame）。在处于核糖体和 mRNA 分子间的中间相位的翻译机构中，每一个密码子被翻译成一个特异氨基酸，它被连续地添加到正在延长的多肽链上。密码子与氨基酸之间的对应由一组称为遗传密码（genetic

code)的规则来决定。除了几个例外情况(见后面),与细胞核为蛋白质编码基因有关的遗传密码是"通用的"(universal),即几乎所有真核生物的细胞核基因和原核生物的基因的翻译,都是由同一组规则所决定的。

通用密码子由表1—2给出。由于1个密码子由3个核苷酸组成,且存在四种不同类型的核苷酸,所以,应有$4^3=64$种可能的密码子。其中61个为特异的氨基酸编码,并被称为有意义密码子(sense codons),其余3个则为无意义密码子或终止密码子(nonsense or stop codons),其作用是充当使翻译过程终止的信号。因为有意义密码子有61个,而蛋白质中的氨基酸却只有20种,所以,大多数氨基酸由1个以上的密码子编码。这样一类密码被称为是简并密码(degenerate code)。编码同一氨基酸的不同密码子称同义密码子(synonymous codons)。相互之间仅在第3位上不同的同义密码子被指定为一个密码子族(codon family)。例如,为缬氨酸编码的4个密码子组成一个四密码子族。与之对比,为丝氨酸编码的6个密码子被分成一个四密码子族和一个二密码子族。

表1—1 基本氨基酸及其三字母和一字母缩写

名 子	三字母缩写	一字母缩写
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic Acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

表1—2 通用密码子

密码子	氨基酸	密码子	氨基酸	密码子	氨基酸	密码子	氨基酸
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

大多数真核生物的蛋白质中第1个氨基酸是甲硫氨酸，由起始密码子(initiation codon)AUG编码。这一氨基酸在成熟蛋白质中通常被移除了。大多数原核生物的基因也用AUG密码子作为起始，起始翻译过程的氨基酸则是甲硫氨酸的衍生物，称甲酰甲硫氨酸。

表1－3　哺乳动物线粒体的密码子*

密码子	氨基酸	密码子	氨基酸	密码子	氨基酸	密码子	氨基酸
UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys
UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys
UUA	Leu	UCA	Ser	UAA	Stop	**UGA**	**Trp**
UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp
CUU	Leu	CCU	Pro	CAU	His	CGU	Arg
CUC	Leu	CCC	Pro	CAC	His	CGC	Arg
CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser
AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
AUA	**Met**	ACA	Thr	AAA	Lys	**AGA**	**Stop**
AUG	Met	ACG	Thr	AAG	Lys	**AGG**	**Stop**
GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly
GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly

* 与通用密码子的差别用黑体字表示。

通用密码子在质体—象维管植物的叶绿体—基因组所采用的独立的翻译过程中也被使用着。比较起来，大多数线粒体基因组和几种细胞核基因组(例如，枝原体属 *Mycoplasma* 和四膜虫属 *Tetrahymena*)则使用与通用密码子不同的密码子。不过，一般说来这些密码子与通用密码子之间只有较小的差别。作为一例，哺乳动物线粒体的密码子列于表1－3。注意，通用遗传密码中编码丝氨酸的密码子有两个被用作终止密码子，而色氨酸和甲硫氨酸都是被两个密码子编码，而不是一个。

1.4　突变

DNA序列在染色体复制过程中，正常情况下被精确地拷贝。然而，极偶然地也会出现错误，从而产生新的序列。这些错误称之为突变(mutations)。突变既可发生在体细胞内又可发生在种系细胞内。由于体细胞突变是不遗传的，所以，在只讨论进化问题时我们将不考虑它们。本书通篇提及的“突变”这个词，将只指在种系细胞中的突变。

突变可以根据受突变事件影响的DNA序列的长短来分类。例如，突变可能影响到一个核苷酸(点突变(point mutations))或几个相互邻近的核苷酸。我们也可按突变事件造成的变化类型，将突变分成：(1)替换(substitutions)，即一个核苷酸被另一个所取代，(2)缺失(deletions)，即从DNA中移去一个或多个核苷酸，(3)插入(insertions)，向序列中添加一个或多个核苷酸，(4)倒位(inversions)，即含有2个或多个碱基对的双链DNA片段转动180°(图1－5)。

核苷酸替换

核苷酸替换分成转换(transition)和颠换(transversions)两类。转换是A和G(嘌呤)之间或T和C(嘧啶)之间的替换。颠换是一个嘌呤和一个嘧啶之间的替换。

发生在蛋白质编码区的核苷酸替换也可通过它们对翻译产物蛋白质的影响来定性。如果替换不引起氨基酸改变(图1－6a)，则被称为是同义的(synonymous)或沉默的(silent)。非同义的或引起氨基酸改变的突变，又可进一步被划分成误义的(missense)和无义的(nonsense)突变。误义突变将受到

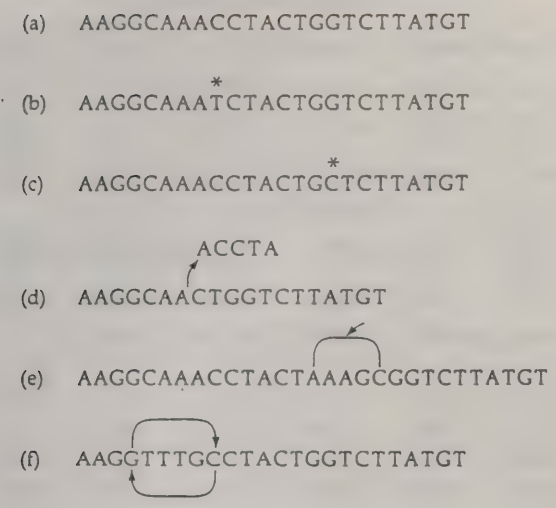


图 1—5　突变的类型

(a)原序列；(b)从 C 到 T 的转换；(c)从 G 到 C 的颠换；(d)缺失序列 ACCTA；(e)插入序列 AAAGC；(f)5′—GCAAAC— 3′ 倒位成 5′— CAAACG— 3′。

影响的密码子变成另一种密码子，后者确定的氨基酸与前者编码的氨基酸互不相同(图 1—6b)。无义突变则将密码子变成一个终止密码子，这样，由于翻译过程在成熟前就结束了，所以最终结果是产生一个截掉尾巴的蛋白质(图 1—6C)。

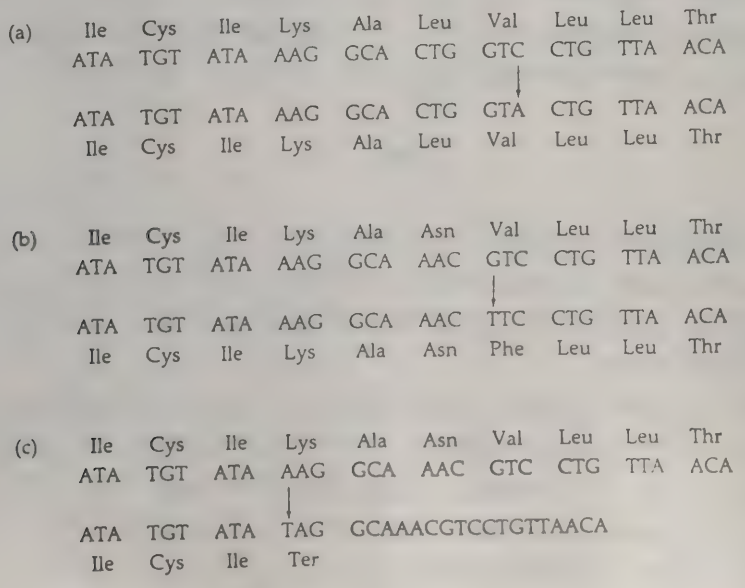


图 1—6　发生在编码区中的替换类型

(a)同义的，(b)误义的，和(c)无义的。

每一个有意义密码子通过一次核苷酸替换可以突变成 9 种其他密码子，例如，CCU(pro)可以经历 6 种非同义的替换，变成 UCU(Ser)，ACU(Thr)，GCU(Ala)，CUU(Leu)，CAU(His)或 CGC(Arg)，以及 3 种同义的替换，变成 CCC，CCA 或 CCG。因为通用遗传密码由 61 个有意义密码子组成，所以有 $61\times9=549$ 种可能的核苷酸替换。如果我们假定核苷酸替换随机地进行且编码区中所有密码子具相同的频率，那么，我们就能根据遗传密码算出不同类型核苷酸替换的期望比例。这些数据

列于表 1—4。由于遗传密码的结构，同义替换主要发生在密码子的第 3 位上。事实上，第 3 位上所有可能发生的改变中，大约 70%是同义的。相比之下，密码子第 2 位上的所有替换都是非同义的，而第 1 位上发生的核苷酸变化，绝大多数也是如此(96%)。

表 1—4　随机蛋白质编码序列中不同类型突变替换的相对频率

替换	数目	百分比
全部密码子总数	549	100
同义的	134	25
非同义的	415	75
误义	392	71
无意义	23	4
第一位总数	183	100
同义的	8	4
非同义的	175	96
误义	166	91
无意义	9	5
第二位总数	183	100
同义的	0	0
非同义的	183	100
误义	176	96
无意义	7	4
第三位总数	183	100
同义的	126	69
非同义的	57	31
误义	50	27
无意义	7	4

图 1—7　不等价交换，结果是子链之一中某一 DNA 序列缺失，而另一子链中则出现同一序列的重复。图中矩形表示某一特定长度的 DNA。

缺失和插入

缺失和插入可能由几种机制造成。机制之一是不等价交换(unequal crossing over)。图 1—7 绘出了一个简单模型，按此方式，两条染色体间不等价交换，结果在一条染色体上某一 DNA 片段缺失，而在另一条上则出现相应的添加。另一个机制是复制滑脱(replication slippage)或滑脱链误配(slipped-strand mispairing)。这类事件发生在含有邻接短重复序列的 DNA 区域中。图 1—8a 表示，在 DNA 复制期间，滑脱可因邻接重复间的误配而引起，而滑脱又可造成某一 DNA 片段缺失或重复的后果，至于究竟是缺失还是重复，则要看滑脱发生在 5′→3′方向还是与之相反的方向。图 1—8b 表明，滑脱误配也可能发生在非复制 DNA 中。造成 DNA 序列插入或缺失的第三种机制是 DNA 转座，这将在第 7 章中加以探讨。

缺失和插入统称裂缝(gaps)，因为当一个带有缺失或插入的序列与原序列比较时，两序列中就将有一个会出现“裂缝”(第 3 章)。裂缝事件中涉及的核苷酸数目，范围从一个或几个到包括成千个核苷

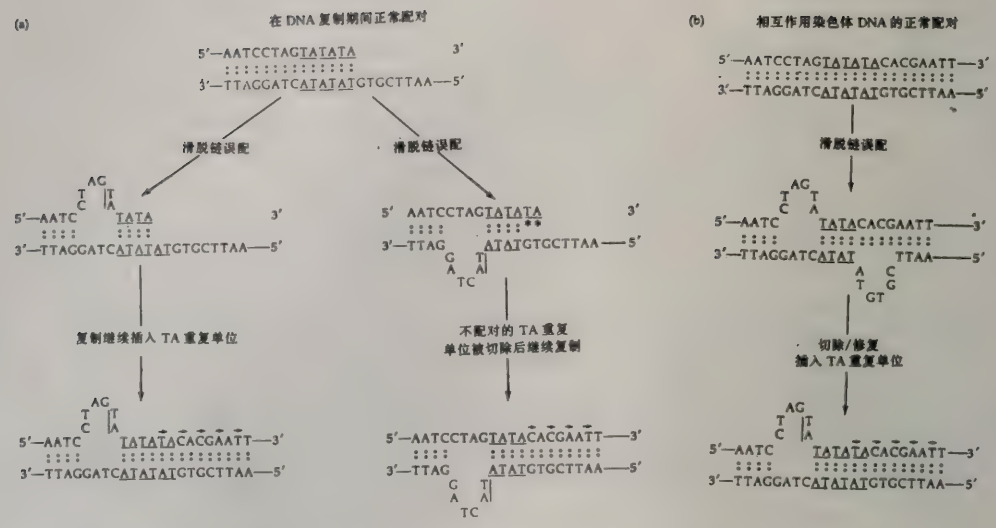


图 1－8　由邻接的重复序列(下面划有横线)间的滑脱误配所产生的重复和缺失。小箭头指示 DNA 合成的方向，圆点表示碱基配对。(a)在 DNA 复制期间 TA 重复中发生的两碱基滑脱。3′→5′方向中的滑脱造成一个 TA 单位插入的结果(左半图)。另一方向中的滑脱则造成一个重复单位缺失的结果(右半图)。右半图中的缺失是正在生长的链的 3′端上，未配对重复单位(标有星号)被切除而造成的，推测系凭借 DNA 聚合酶的 3′→5′外切酶活性。(b)在非复制 DNA 中 TA 重复中发生的两碱基滑脱。误配区形成单链环，它们可能是切离和误配修复的目标。结果(是缺失还是插入)依赖于倒底是哪条链被切离和被修复，以及在 DNA 修复过程中究竟是哪条链被当作模板。自 Levinson 和 Gutman (1987)修改而成。

酸的连续长段。裂缝的长度基本上展示出双峰型频率分布，短裂缝(到 20－30 核苷酸为止)常常是由 DNA 复制过程中的错误所造成的，象上面讨论过的滑脱链误配，而长段的插入和缺失则主要由不等价交换或 DNA 转座所造成。

在编码区，如果裂缝事件所涉及的核苷酸数不是 3 的倍数，则将造成阅读框架的移动，于是裂缝下游的编码序列将会按错误的相位阅读。这种突变称框架移动(frameshift)突变。结果，一个裂缝不仅会导入许多氨基酸的改变，而且还可能使终止密码子被抹掉或者在相位中导入新的终止密码子，后果是产生长度异常的蛋白质(图 1－9)。

突变的空间分布

突变不是在整个基因组中随机出现的。有些区域比另一些区域更容易突变，它们被称为突变的热点(hotspots)。有一个突变热点是二核苷酸 5′－CG－3′，其中胞嘧啶常常被甲基化并被错误地复制，最后它就变成为 5′－TG－3′。二核苷酸 5′－TT－3′在原核生物中是一个突变热点，但在真核生物中却通常不是。在细菌中，含有短回文(palindromes)(即，两条互补链读起来相同的序列，如 5′－GCCG-GC－3′，5′－GGCGCC－3′和 5′－GGGCCC－3′)的 DNA 内的区域，被发现比其他区域更易于突变。在真核生物的基因组中，串接重复常常是缺失和插入的热点，或许这是滑脱链误配所造成的结果。

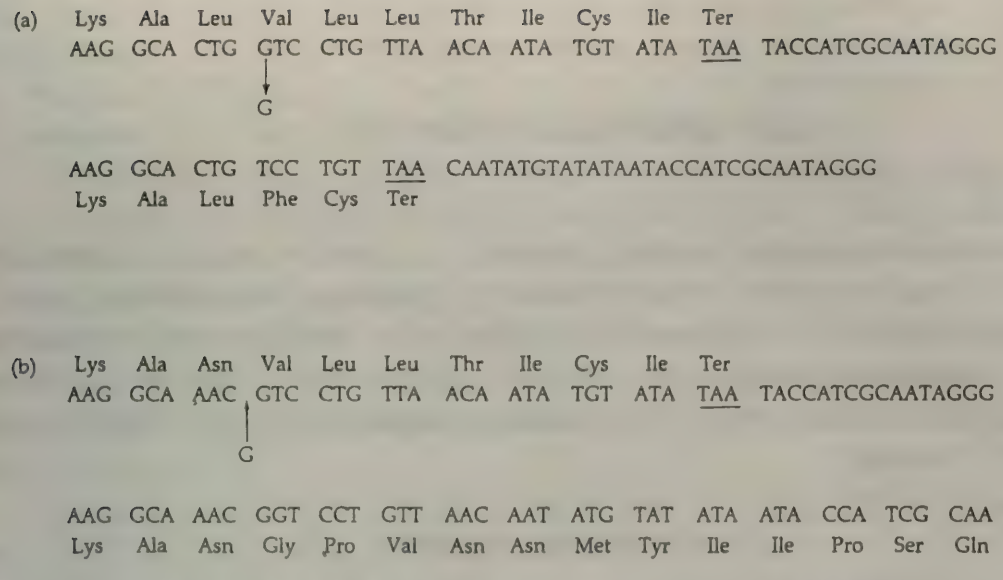


图 1－9　由缺失或插入造成的阅读框架发生框架移动的例子。(a)缺失一个 G 造成成熟前终止。(b)插入一个 G，结果抹掉了一个终止密码子。终止密码子下划有横线。

习题

1. 从某种哺乳动物中找出两个完全定序了的为蛋白质编码的基因(你可以应用数据库或一本杂志)。确定它们每一个的内含子和编码外显子间的长度比。用取自果蝇的两个为蛋白质编码的基因再做上述处理。哪些基因的内含子与外显子长度比更大?

2. 从某种哺乳动物中找出一个完整的 mRNA 序列。删去你在编码区中最先碰到的两个核苷酸 CT。判断是否有一个成熟前终止密码子进入阅读框架，或者是否翻译会超越原终止密码子而延长。

3. 用习题 2 中的序列，在编码区中第 6 个 C 后插入一个 A。插入将会怎样影响翻译?

4. 用习题 2 中的序列，判断有多少密码子能经一次核苷酸替换而突变成终止密码子。

5. 用习题 2 中的序列，判断密码子第 3 位中有多少转换是同义的，而又有多少颠换是同义的。

后继阅读文献

Darnell, J. E., H. F. Lodish and D. Baltimore. 1990. *Molecular Cell Biology*, 2nd Ed. Scientific American Books, New York.

Lewin, B. 1990. *Genes* Ⅳ. Oxford Universtity Press, New York.

Stryer, L. 1988. *Biochemistry*, 3rd Ed. Freeman, New York.

Suzuki, D. T., A. J. F. Griffiths, J. H. Miller and R. C. Lewontin. 1989. *An Introducition to Genetic Analysis*, 4th Ed. Freeman, New York.

Watson, J. D. N. H. Hopkins, J. W. Roberts, J. A. Steitz and A. M. Weiner. 1987. *Molecular Biology of the Gene*, 4th Ed. Benjamin/Cummings, Menlo Park, CA.

2 群体中基因的动力学

群体遗传学探讨发生在群体内的遗传变化。本章将介绍某些对理解分子进化来说至关重要的、群体遗传学的基本原理。群体遗传学的一个基本问题，是确定一个突变基因的频率在各种进化力量的影响下，将如何随时间而变化。此外，从长期的观点看，重要的是决定一个新的突变变异型完全替代群体中的老变异型的概率，以及估计替代过程将会进行得多快。与形态上的改变不同，许多分子变化好象对生物的表型只有很小的影响，所以，分子变异型的频率受机遇的影响强烈。因此，在处理分子进化时，机遇因素应该被考虑进去。

2.1 等位基因频率方面的改变

一个基因在染色体或基因组上的位置称基因座位(locus)，某一给定基因座位上所选取的基因形式称等位基因(alleles)。在一个群体中，某一基因座位上可能有一个以上的等位基因存在，它们的相对比例称等位基因频率(allele frequencies)或基因频率(gene frequencies)。例如，我们假定，一个大小为N的单倍体群体中，某一基因座位上有各具 n_1 和 n_2 个拷贝的两个等位基因。那么，它们的等位基因频率就分别等于 $\frac{n_1}{N}$ 和 $\frac{n_2}{N}$。注意，其中 $n_1+n_2=N$，而 $\frac{n_1}{N}+\frac{n_2}{N}=1$。

进化是在群体的遗传组成方面发生变化的过程。因此，进化过程中最基本的部分是等位基因频率随时间的变化。事实上，从进化的观点看，一个新突变要变得有意义就必须增加它自己的频率，并最终在群体中被固定(fixed)(即在随后的世代中所有个体将共有同一种突变型等位基因)。如果不增加自己的频率，那么这个突变将对该物种的进化史几乎没有影响。对于一个要增加频率的突变型等位基因来说，必须是某些因子而不是突变来掺入作用。这些因子包括自然选择，随机遗传漂变，重组和迁徙。

为了认识进化的过程，我们必须研究以上因子是如何影响等位基因频率的变化的。本书里，我们只讨论自然选择和随机遗传漂变。在涉及形态学性状的经典进化研究中，自然选择被看成是进化的主要驱动力量。与此成鲜明对比的是，在分子水平上随机遗传漂变被认为在进化中起重要作用。

研究群体中的遗传变化有两种数学途径：决定性的和随机性的。决定性模型(deterministic model)较简单。它假定，群体中等位基因频率从这一代到下一代的改变是以唯一的方式发生的，并且能根据关于初始条件的知识毫不含糊地预测出来。严格说来，此途径仅当两个条件满足时才能应用：(1)群体在大小上是无限的，和(2)环境或者不随时间而变或者按决定性规则而变。这些条件在自然界显然绝不会得到满足，因此，纯决定性途径对于描述群体中等位基因频率随时间的变化可能是不够的。随机的或不可预料的等位基因频率方面的波动也必须被考虑进去。

处理随机波动需要一种不同的数学途径。随机性模型(stochastic models)假定等位基因频率的变化按或然性方式发生，即，从这一世代中关于条件的知识我们不能含糊地预料下一世代里的等位基因频率，而只能决定可能出现的某些等位基因频率的出现概率。显然，随机性模型比决定性模型更可取，因为它们建立在更为现实的假定之上。不过，决定性模型在数学处理上要容易得多，而且在某些条件下它们也能得出足够精确的近似结果。以下，我们将按决定性方式处理自然选择。

2.2 自然选择

自然选择(natural selection)被定义成：一个群体内遗传上不同的个体或基因型的有差别的增殖。

有差别的增殖是由个体间如死亡率、能育性、生殖力、交配成功和后代生活力等因子方面的差异所造成的。自然选择是个体间在与增殖有关的性状方面存在遗传变异的必然结果。如果一个群体由在这类性状方面相互无差别的个体所组成，则它将不会受到自然选择。选择导致等位基因频率随时间而变。然而，仅仅是等位基因频率从一代到另一代发生变化，并不一定表示自然选择在起作用。别的过程，例如随机遗传漂变，也能导至等位基因频率随时间的改变(见后面的内容)。

一个基因型的适合度(fitness)，通常用 w 表示，是一个关于该个体的生存和增殖能力的尺度。不过，由于一个群体的大小通常受其所处环境的负载容量限制，所以，某一个体的进化成功不是由其绝对适合度(absolute fitness) 而是由其与群体中其他基因型相比的相对适合度(relative fitness)所决定的。在自然界，不能预期一个基因型的适合度在所有世代和所有环境条件下保持不变。然而，对每种基因型指定一个恒定的适合度值，我们就可以得出一些简单的理论公式，它们对理解由自然选择导致的群体遗传结构变化的动力学是很有用的。在某些最简单的模型里，我们假定生物的适合度仅由其遗传组成所决定。我们还假定，所有基因座位对个体的适合度的贡献相互独立，所以每一基因座位都能被隔离开来处理。

群体中产生的大多数新突变型都会降低其携带者的适合度。这类突变将受到淘汰性选择并且最终将会被从群体中去除。这种类型的选择称负选择或纯洁化选择(negative or purifying selection)。偶尔，某一新突变可能与群体中最好的等位基因一样合适，这样的突变即为选择中性的(neutral)，且其命运将不由选择所决定。极罕见地，可能会产生一个能给其携带者以选择优势的突变型。这类突变将受到正选择或有利选择(positive or advantageous selection)。如果这一新突变型仅在杂合子情况下有利，在纯合子情况下则无优势，那么结果选择体制将为超显性选择(overdominant selection)。

下面我们将考虑一个有两个等位基因，A_1 和 A_2 的基因座位的例子。每一等位基因可被指定一个内在适合度值；它可能是优势的、劣势的，或中性的。然而，这种指定只适用于单倍体生物。在两倍体生物中，适合度最后由该基因座位上的两个等位基因间的相互作用所决定。两个等位基因，将有三种可能的两倍体基因型：A_1A_1，A_1A_2 和 A_2A_2，它们的适合度则可分别用 w_{11}，w_{12}和 w_{22}来表示。

给定某一群体中等位基因 A_1 的频率为 p，该互补等位基因 A_2 的频率为 $q=1-p$，我们可以证明，在随机交配下 A_1A_1，A_1A_2 和 A_2A_2 的频率将分别为 p^2，2pq 和 q^2。一个维持着这种基因型比的群体，被认为是处于哈迪一温伯格平衡(Hardy-Weinberg equilibrium)。注意，$p=p^2+\frac{1}{2}(2pq)$以及 $q=\frac{1}{2}(2pq)+q^2$。

在一般情况下，三种基因型被指定为以下适合度值和初始频率：

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	w_{11}	w_{12}	w_{22}
频率	p_2	2pq	q^2

现在让我们考虑一下等位基因频率随选择而改变的动力学。假定三种基因型的频率及其适合度如上，则三种基因型 A_1A_1，A_1A_2 和 A_2A_2 对下一世代的相对贡献，将分别为 p^2w_{11}，$2pqw_{12}$和 q^2w_{22}。因此，等位基因 A_2 的频率在下一世代将变成

$$q' = \frac{pqw_{12} + q^2w_{22}}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \tag{2.1}$$

每世代等位基因 A_2 的频率改变量被表示成 $\Delta q=q'-q$。我们可以证明

$$\Delta q = \frac{pq[p(w_{12} - w_{11}) + q(w_{22} - w_{12})]}{p^2w_{11} + 2pqw_{12} + q^2w_{22}} \tag{2.2}$$

下面，我们假定 A_1 是该群体中的原等位基因或"老"等位基因。然后，我们来考虑一个新突变型等位基因 A_2 出现后，等位基因频率变化的动力学。为了数学上的便利，我们给予 A_1A_1 基因型的适合度值为 1。新产生的基因型 A_1A_2 和 A_2A_2 的适合度将有赖于 A_1 和 A_2 间相互作用的模式。例如，如果 A_2 对 A_1 是完全显性的，那么 w_{11}，w_{12}和 w_{22}可分别写成 1，1+s，和 1+s，如果 A_2 完全隐性，则适合度分别为 1，1，1+s。这里 s 是带有 A_2 的基因型与 A_1A_1 间的适合度差异。s 的值为正表示与 A_1A_1 比较

适合度增高，而为负时则表示适合度降低。

共显性

在选择的共显性模式(codominant mode of selection) 或基因选择(genic selection)中，两种纯合子有着不同的适合度值，而杂合子的适合度则为两种纯合基因型的适合度的平均值。三种基因型的相对适合度值可写成：

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	1	1+s	1+2s

根据等式 2.2，在共显性下我们得等位基因 A_2 的每世代频率改变如下：

$$\Delta q = \frac{spq}{1 + 2spq + 2sq^2} \tag{2.3}$$

图 2—1 显示具 s=0.01 的等位基因 A_2 的频率增加情况。我们看到，共显性选择总是在消耗别的等位基因频率的前提下增加其中一种等位基因的频率，而与等位基因在群体中的相对频率无关。因此，基因选择是一种定向选择(directional selection)。注意，在低频率下对一个共显性等位基因的选择

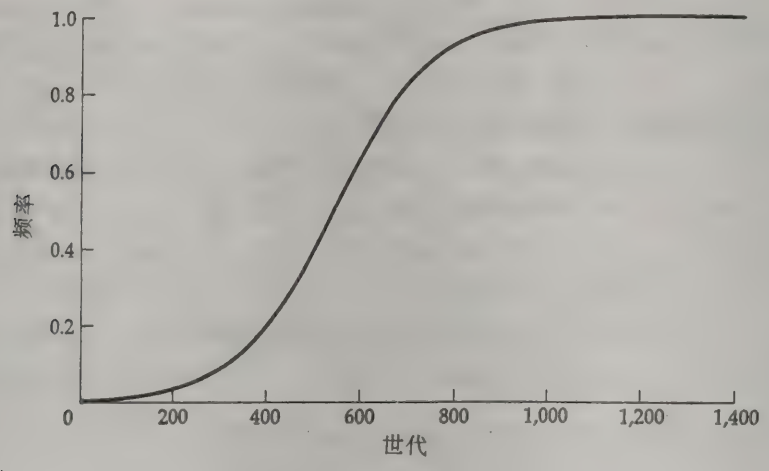


图 2—1　具 S=0.01 的共显性优势等位基因，在其于第 0 代作为突变的结果而出现后的频率。

不是十分有效(即，等位基因频率的变化缓 慢)。原因在于，当 A_2 频率低时，等位基因 A_2 处于杂合子中的比例较大。例如，当 A_2 的频率为 0.5 时，50% 的 A_2 等位基因由杂合子携带，而若 A_2 的频率为 0.01，则 99% 的这种等位基因都处于杂合子中。因为杂合子带有两种等位基因，受到的选择压力比纯合子 A_2A_2 的弱(即 s 和 2s 的关系)，所以在低 q 值下等位基因频率的总改变量就小。

超显性

在选择的超显性模式(overdominant mode of selection)，杂合子有最高的适合度。于是：

基因型	A_1A_1	A_1A_2	A_2A_2
适合度	1	1+s	1+t

这里 s>0 且 s>t。根据 A_2A_2 的适合度是高于、等于或低于 A_1A_1 的适合度，t 可相应地取正值、零、或负值。等位基因频率的变化表示成

$$\Delta q = \frac{pq(2sq - tq - s)}{1 + 2spq + tq^2} \tag{2.4}$$

图 2—2 表示一个受超显性选择的等位基因的频率变化。与造成一个等位基因最终从群体中消失的共显性选择体制不同，在超显性选择下，群体迟早将达到一种两等位基因共存的平衡状态。平衡达到后将看不到等位基因频率的进一步变化(即 $\Delta q=0$)。所以，超显性选择属于一种称为平衡选择或稳定化选择(balancing or stabilizing selection)的选择体制。

处于平衡时的等位基因 A_2 的频率可通过令等式 2.4 的 $\Delta q=0$ 而解出：

$$\hat{q}=\frac{s}{2s-t} \tag{2.5}$$

当 t=0(即两种纯合子有相同的适合度值)时，两种等位基因的平衡频率将都是 50%。

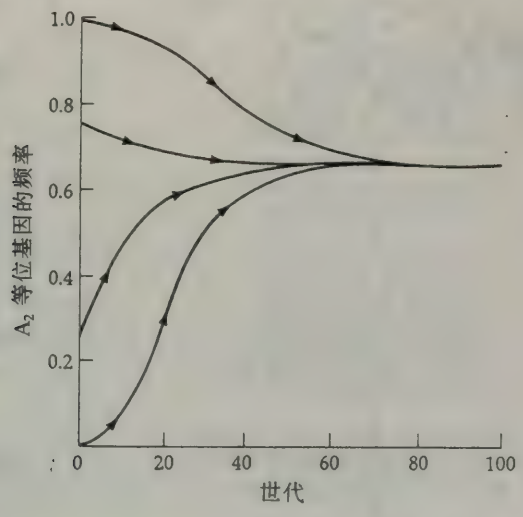


图 2－2　一个受超显性选择的等位基因的频率变化。起始频率从上到下各为：0.99，0.75，0.25 和 0.01；s=0.250 及 t=0.125。因为 s 和 t 的值都特别大，所以等位基因频率的变化迅速。注意，在 q=0.667 处有一个稳定平衡。自 Hartl 和 Clark(1989)修改而成。

2.3　随机遗传漂变

正如上面提到的，自然选择不是引起等位基因改变的唯一因子。等位基因频率变化也可因机遇而产生，虽然在这种情况下变化不是有方向性的而是随机的。产生等位基因频率的随机波动的一个重要因子是生殖过程中配子的随机取样(图 2－3)。之所以出现取样问题，是因为在自然界中绝大多样情况下，任一个世代中可用配子的数目都远大于下一世代所产生的成年个体数。换句话说，只有一小部分配子成功地发育到成体。在一个经受孟德尔式分离的两倍体群体中，即使没有配子过剩，即每一个体对下一世代正好贡献两个配子，取样问题也会发生。原因是，杂合子能产生两种配子，而传给下一世代的两个配子由于机遇却可能是同一类型的。

为了看取样造成的后果，让我们考虑一个群体中所有个体有同样的适合度以至于选择不起作用的理想情形。我们再把问题进一步简化，考虑一个没有重叠世代的群体(即一群同时繁殖的个体)，这样，任一给定世代都能与前一世代和后一世代毫不含糊地区分开来。该被考虑的群体为两倍体，由 N 个个体所组成，所以，对任一给定的基因座位来说该群体含有 2N 个基因。接下来我们处理具有两个等位基因 A_1 和 A_2、频率分别为 p 和 q=1－p 的一个基因座位的简单例子。当从无限的配子库中抽取 2N 个配子时，样本中正好含有 i 个 A_1 型基因的概率 p_i 由二项式概率函数给出：

$$p_i=\frac{(2N)!}{i!(2N-i)!}p^iq^{2N-i} \tag{2.6}$$

对于两种等位基因共存(即 $0<p<1$)的群体而言，p_i 总是大于零的，所以，等位基因频率可以无需选择的帮助而一代一代地改变。

只是由于机遇而造成等位基因频率变化的过程称随机遗传漂变(random genetic drift)。不过要注意，随机遗传漂变也可由不是配子取样的过程所引起。例如，选择强度上的随机变化也能导至等位基因频率方面的随机变化。

我们用图 2－4 来说明随机取样对不同大小群体的等位基因频率的影响。等位基因频率一代一代地改变，但变化的方向在任一时间点上都是随机的。随机漂变的最明显特征是，等位基因频率的波动

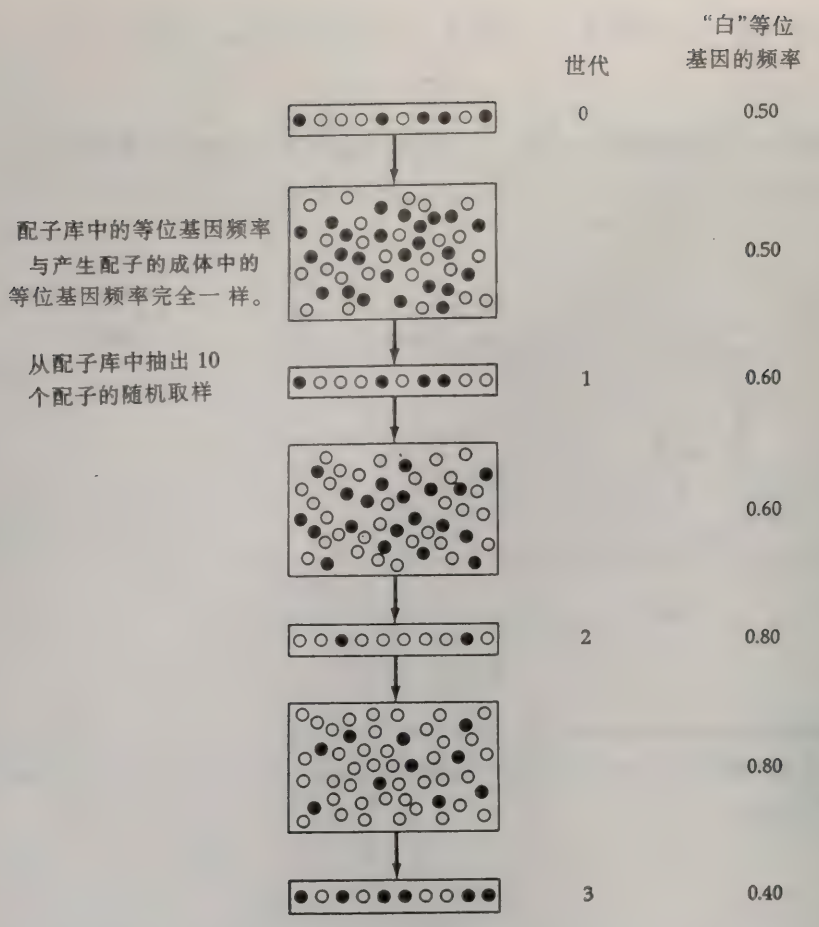


图 2—3　配子的随机取样。假定每一世代的配子库(大长方块)中的等位基因频率精确地反映了亲本世代(小长方块)的成体中的等位基因频率。因为群体大小是有限的,所以等位基因频率上下波动。自 Bodmer 和 Cavalli-Sforza (1976)修改而成。

在小群体中比在大群体中要突出得多。

让我们跟踪由相继世代中随机遗传漂变过程造成的等位基因频率改变的动力学。等位基因 A_1 的频率被写成 p_0、p_1、p_2、…、p_t,其中下标指世代数。等位基因 A_1 的起始频率为 p_0。类似地,等位基因 A_2 的频率为 q_0、q_1、q_2、…、q_t。在没有选择时,我们期望 p_1 与 p_0 相等,q_1 与 q_0 相等,且对以后世代都是如此。然而,群体是有限的这一事实,则意味着 p_1 只是在平均上(即重复取样无限次时)才等于 p_0。而实际情况是,取样在每一世代里只发生一次,所以 p_1 通常不同于 p_0。在第 2 代里,A_1 的频率(p_2)将不再依赖于 p_0 而依赖于 p_1。类似地,在第 3 代 A_1 的频率(p_3)将既不依赖于 p_1 也不依赖于 p_0,而是依赖于 p_2。所以,随机遗传漂变的最重要的性质是其累积行为(cumulative behavior),即随着一代代地传下去,某一等位基因的频率将倾向于越来越偏离其起始频率。

为了看一看随机遗传漂变的累积效应,让我们考虑一下下面的数字例子。某一群体由 5 个两倍体个体所组成,在某一基因座位上的两个等位基因 A_1 和 A_2,频率均为 50%。现在我们要问,"在下一世代得到同样的等位基因频率的概率是多少?"应用等式 2.6,我们得到值为 0.25 的概率。换句话说,在 75%的情况下第 2 代的等位基因频率将不同于起始频率。而且,在以后的世代里保留着起始等位基因频率的概率也不再是 0.25,而将会越来越小。例如,在第 3 代群体中有 50%A_1 的概率约为 18%,而 10 代以后该概率就只有约 5%(图 2—5)。相应地,A_1 或者 A_2 丢失的概率则随时间的推移而增加,因为,所有被选取的配子正好带有同样的等位基因的事件,在每一世代中都具有一个有限的概率。一旦某一种等位基因的频率达到 0 或者 1,它的频率在以后的世代里就不再变化了。第一种情形被称为丢失(lose)或灭绝(extinction),第二种则称之为固定(fixation)。如果取样过程持续一个相当长的时期,则这类可能情形将必然会达到。

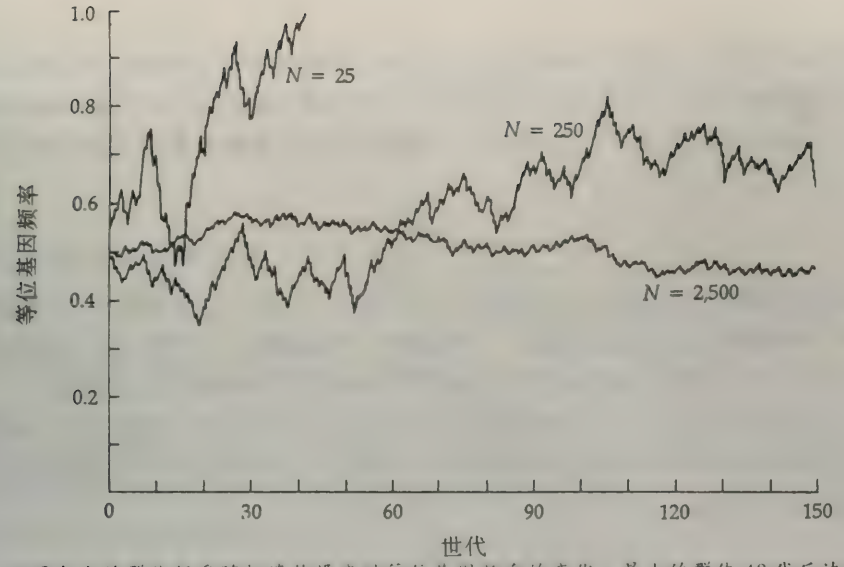


图 2—4　不同大小的群体经受随机遗传漂变时等位基因频率的变化。最小的群体 42 代后达到固定。另两个群体 150 代以后还是多态性的，但如果实验继续得足够长则终将会达到固定(等位基因频率＝0%或 100%)。自 Bodmer 和 Cavalli-Sforza (1976)。

随机遗传漂变的最终结果是一个等位基因的固定以及所有别的基因的丢失。要不发生这样的结果，除非通过象突变或迁移这样的过程使等位基因不断地向群体中输入，或者由某种平衡选择积极地维持着多态现象。

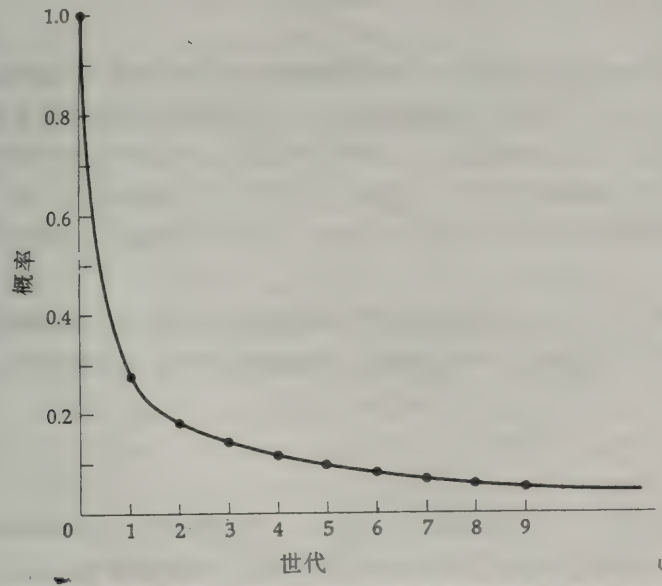


图 2—5　两个选择中性的等位基因，有 N　) 和 p　0..)，其维持与起始等位基因频率相同的概率随时间而变的情形。

2.4　有效群体大小

群体生物学的一个基本参数是群体大小(population size)N，它被定义成：一个群体中的总个体数。然而，从群体遗传学和进化的观点看，将被考虑的有关数仅由那些积极参入繁殖过程的个体所构成。由于并非所有个体都参入繁殖，所以，影响进化过程的群体大小与调查统计的大小是不同的。群体的这部分称有效群体大小(effective population size)，并用 Ne 来表示。

一般来说，Ne 小于 N，有时甚至比 N 要小得多。有各种各样的因素能造成这种差异。例如，在一个具有重叠世代的群体中，在任何时间里，群体的部分将由处于生殖前或生殖后阶段的个体所组成。由于发育阶段重叠，有效大小可能比调查统计大小要小很多。例如，据根井和今泉(Nei 和 Imaizumi，1966)报导，人类的 Ne 仅略大于 N/3。

有效群体大小比调查统计大小要小，也可能发生在涉及生殖的雄性个体数不等于雌性个体数的时候。这种情况在多配偶制的物种，象社会性哺乳类和领土性鸟类，或者存在无生殖阶层的物种（例如蜜蜂、蚂蚁和白蚁）中尤其明显。如果一个群体由 N_m 个雄体和 N_f 个雌体所组成，则 Ne 将由以下公式给出：

$$Ne = \frac{4N_mN_f}{N_m + N_f} \tag{2.7}$$

注意，除非雌性个体数与雄性个体数相等，否则 Ne 总是要比 N 小。举一个极端的例子，我们假定一个大小为 N 的群体中，所有雌体（N/2）以及仅有一个雄体参入生殖过程。应用等式 2.7，我们看到 Ne=2N/(1+N/2)。如果 N 比 1 大得多，则 Ne 将等于 4，而与调查统计到的群体大小无关。

有效群体大小也可能因群体大小的长期变动而大为降低，后者可由象环境的灾变，生殖的循环模式地区性灭绝和重建事件等因素所引起。例如，在一个经历 n 个世代的物种中，长期有效群体大小（long-term effective population size）由：

$$Ne = \frac{n}{\frac{1}{N_1} + \frac{1}{N_2} + \cdots + \frac{1}{N_n}} \tag{2.8}$$

给出，其中 Ni 是第 i 世代的群体大小。换言之，Ne 等于 Ni 值的调和平均，结果它更接近于 Ni 中的最小值而不是最大值。类似地，如果一个群体经历一次瓶颈，则有效群体大小将大为降低。

2.5 基因替换

基因替换（gene substitution）被定义成一个过程，经此，一个突变型等位基因将完全代替群体中占优势的或"野生型"（wild type）等位基因。在此过程中，一个突变型等位基因在群体中产生，通常以一个拷贝出现，并且经一定世代数以后被固定。一个新的等位基因固定所花的时间称固定时间（fixation time）。不过，并非所有新突变型都能达到固定。事实上，它们中的大多数在几个世代后就会被丢失。所以，我们还需要谈谈固定概率（fixation probability）问题，以及讨论影响着一个新突变型等位基因在群体中固定的机会的因素。

新突变在群体内不断地产生着。结果，基因替换不断地发生，用一个等位基因替代另一个，而有时它自己又被一个新的等位基因所替代。所以，我们才能谈到基因替换的速率，即每单位时间的替换或固定数。

固定概率

某一特定等位基因在群体中被固定的概率有赖于：(1)起始频率，(2)选择优势度或劣势度 s，(3)有效群体大小 Ne。下面，我们将考虑基因选择的情形，并假定三种基因型 A_1A_1，A_1A_2 和 A_2A_2 的相对适合度分别为 1，1+s 和 1+2s。

木村（Kimura，1962）曾证明，A_2 的固定概率由

$$p = \frac{1 - e^{-4N_esq}}{1 - e^{-4N_es}} \tag{2.9}$$

给出，其中 q 为等位基因 A_2 的起始频率。由于当 x 很小时 $e^{-x} \approx 1-x$，所以等式 2.9 在 x 趋近于 0 时简化成 $p \approx q$。所以，对于一个中性等位基因而言，固定概率等于它在群体中的频率。例如，一个频率为 40% 的中性等位基因，它将在 40% 的事例中固定而在 60% 的事例中丢失。这可以从直观上认识，因为在中性等位基因的场合，固定是由不偏袒任何等位基因的随机遗传漂变所造成的。

我们注意到，一个新突变型在大小为 N 的两倍体群体中以一个拷贝出现，它将有 1/(2N) 的起始频率。一个突变型等位基因的固定概率 p，可在等式 2.9 中用 1/(2N) 代替 q 来得出。当 $s \neq 0$ 时，

$$p = \frac{1 - e^{-\frac{2Nes}{N}}}{1 - e^{-4Nes}} \tag{2.10}$$

对于一个中性突变，等式 2.10 变为

$$p = \frac{1}{2N} \tag{2.11}$$

如果群体大小与有效群体大小相同，则等式 2.10 简化成

$$p = \frac{1 - e^{-2s}}{1 - e^{-4Ns}} \tag{2.12}$$

如果 s 的绝对值较小，则我们得

$$p \approx \frac{2s}{1 - e^{-4Ns}} \tag{2.13}$$

对于 s 取正值而 N 的值较大时，等式 2.13 简化为

$$p \approx 2s \tag{2.14}$$

所以，如果一个有利突变产生于一个大群体中且其超越其他等位基因的选择优势度较小，比如说不到 5%，那么其固定的概率将近似等于其选择优势度的两倍。例如，若一个具 s=0.01 的新突变在群体中产生，则其最终固定的概率即为 2%。

现在我们考虑一个数学例子。一个新的突变型在一个有 1000 个个体的群体中产生。如果(1)它是中性的，(2)它给出 0.01 的选择优势度，或(3)它具有 0.001 的选择劣势度，那么该等位基因将在群体中固定的概率各为多少？为了使问题简化，我们假定 N=Ne。对中性场合来说，固定的概率是 1/(2N)=0.05%。由等式 2.12，我们得到优势和劣势情况下固定的概率各为 2%和 0.004%。这些结果是很值得注意的，因为它们从本质表示一个有利突变并不总是会在群体中固定。事实上，具有 s=0.01 的选择优势度的所有突变中，98%将会因机遇而丢失。此理论结果极为重要，因为它表明，把适应性进化当成有利突变从群体中产生并总是会在以后的世代中兴旺的过程是幼稚的观点。而且，即使是轻微有害的突变也具有一个在群体中固定的有限概率，尽管这是一个小概率。一个有害的等位基因也可能会在群体中固定，仅仅这一事实就能有力地说明机遇作用在进化期间对决定突变的命运方面的重要性。

固定时间

一个等位基因固定或丢失所需要的时间与该等位基因的频率和群体的大小有关。达到固定或丢失的平均时间随着等位基因的频率分别趋近于 1 或 0 而越来越短。

在处理新突变时，对固定和丢失分开处理要方便一些。以下，我们对那些最终将在群体中固定的突变型的平均固定时间进行处理。这一变量称为条件固定时间(conditional fixation time)。在一个两倍体群体中一个新突变的起始频率，据定义为 q=1/(2N)，此例的平均条件固定时间 t，已由木村和太田(Kimura 和 Ohta，1969)算出。对于一个中性突变而言，它由下式所近似给出：

$$\bar{t} = 4N \text{ 代} \tag{2.15}$$

而对于一个具选择优势度 s 的突变来说，它近似为

$$\bar{t} = (2/s)ln(2N) \text{ 代} \tag{2.16}$$

这里 ln 表示自然对数函数。

为了说明不同突变类型间的差别，我们假定，某种哺乳动物具有约 10^6 的有效群体大小和 2 年的世代(间隔)时间。在这些条件下，一个中性突变在群体中固定平均要花 8 百万年。相比之下，一个具 1%的选择优势度的突变在同样的群体中固定则只花约 5800 年。有趣的是，一个具选择劣势度 s 的有害等位基因其条件固定时间与具选择优势度 s 的有利等位基因的正好相等(Maruyama 和 Kimura，1974)。如果一个有害等位基因被给予高丢失概率，则这一点就可从直观上加以理解。即，对于一个注定要在群体中固定的有害等位基因而言，固定必须发生得非常快。

图 2－6 中我们给出了关于有利的和中性的突变的基因替换动力学过程。我们注意到，有利突变或者迅速丢失或者迅速在群体中固定。与之不同的是，中性等位基因的频率变化则是缓慢的，且固定时间比有利突变型的要长得多。

基因替换速率

我们现在考虑一下替换速率(rate of substitution)，它被定义为：每单位时间里达到固定的突变型

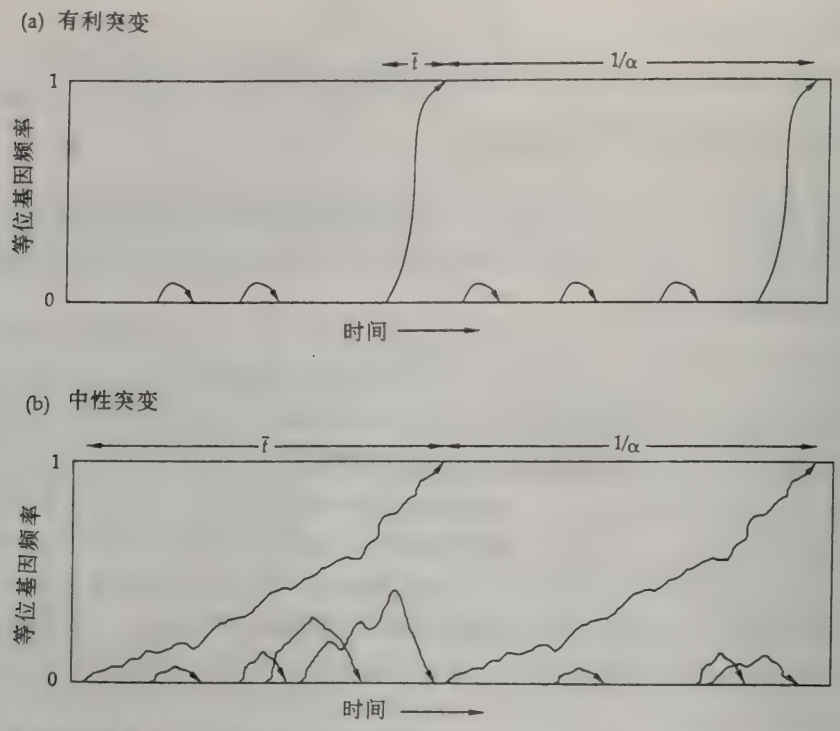


图 2—6 基因替换的动力学过程(a)有利的和(b)中性的突变。有利突变或者很快从群体中丢失或者迅速被固定，所以它们对遗传多态性的贡献小。另一方面，中性等位基因的频率相比之下变化非常缓慢，所以产生大量的瞬时多态。$\bar{t}$ 为条件固定时间，而 $1/\alpha$ 是接连的两固定事件间的平均时间。自 Nei(1987)。

的数目。我们 将先考虑中性突变。如果突变以每世代每基因 u 的速率发生，那么在一个大小为 N 的两倍体群体中产生的突变型数即为每世代 2Nu。因为这些突变中每一个的固定概率都是 1/(2N)，所以，我们用总突变数乘以它们的固定概率即可得到中性等位基因的替换速率：

$$K = u \tag{2.17}$$

于是，对中性突变来说，替换速率等于突变率，这是一个惊人的简单的结果(Kimura，1968a)。如果我们注意到，在一个大群体中，每代产生的突变数虽高但每个突变的固定概率却低，则该结果就可以直观地加以理解。相比之下，在一个小群体中每代产生的突变数虽低，但每一突变的固定概率却高。结果是，中性突变的替换速率与群体大小无关。

对有利突变而言，替换速率可用突变率乘以如等式 2.14 给出的每基因的固定概率来求得。对于具 s>0 的基因选择，我们得：

$$K = 4Nsu \tag{2.18}$$

换言之，对基因选择的情况，替换速率有赖于群体大小(N)、选择优势度(s)、以及突变率(u)。

2.6 遗传多态性

如果一个基因座位在群体中有两个或更多的等位基因存在，则它就被说成是多态性的(polymorphic)。然而，如果其中一个等位基因有非常高的频率，比如说 99%或更高，那么，其他等位基因看来都不能在样本中被观察到，除非样本特别大。所以，从实际出发，如果那种最常见的等位基因的频率小于 99%，则该基因座位通常被定义为多态性的。该定义显然有点武断，而且在文献中还可发现有人用别的标准。

测度某一群体的多态性程度的最简单方法之一，是通过将多态性基因座位数用所取样本的基因座位总数来除，以算出多态性基因座位的平均比例。然而，该测度与被研究的个体数有关。群体内遗传变异性的更为合适的尺度是平均期望杂合度(expected heterozygosity)或基因多样度(gene diversity)。该尺度不依赖多态性的武断描绘，可以从有关基因频率的知识中直接算出，并且受取样作用的影响较小。一个基因座位上的基因多样度定义为：

$$h = 1 - \sum_{i=1}^{m} x_i^2 \tag{2.19}$$

这里 x_i 是等位基因 i 的频率而 m 为该基因座位上等位基因的总数。对任一给定的基因座位而言，h 是随机地从该群体中选取的两个等位基因互不相同的概率。所有被研究过的基因座位的 h 值的平均值，H，可用作群体内遗传变异性程度的一种估计。

基因多样度尺度 h 和 H，曾被广泛用于电泳数据和限制酶数据。不过，它们可能不适用于 DNA 顺序数据，因为自然界中 DNA 水平上的遗传变异程度是相当高的。特别地，当所考虑的序列很长时，样本中的每一序列看来都与别的序列有一个或多个核苷酸的差别，在大多数情况下 h 和 H 都将接近于 1。所以，这些基因-多样度尺度将无法区别不同的基因座位或群体，也不再是多态性的信息尺度。

对于 DNA 顺序数据，关于群体中多态性的更合适的尺度，是任意两个随机选取的序列间每位点的平均核苷酸差异数。该尺度称核苷酸多样度(nucleotide diversity)(Nei 和 Li，1979)，用 π 表示：

$$\pi = \sum_{ij} x_i x_j \pi_{ij} \tag{2.20}$$

这里 x_i 和 x_j 分别为 DNA 序列的第 i 种和第 j 种类型的频率，而 π_{ij} 是 DNA 序列的第 i 种和第 j 种类型间不同核苷酸的比值。

目前，关于 DNA 序列水平上的核苷酸多样度，仅有为数不多的一些研究。这类研究之一涉及果蝇 *D. melanogaster* 的乙醇脱氢酶(Adh)基因座位。跨越 Adh 区域的 11 个序列已由克赖特曼(Kreitman，1983)定序。这些线性排列的序列长达 2379 个核苷酸。如果不管缺失和插入，则有 9 个不同的等位基因，其中一个由 3 个序列表示，而其余的则各由一个序列表示(图 2－7)。所以，频率 x_1—x_8 各为 1/11，而频率 x_9 为 3/11。43 个核苷酸位点是多态性的。首先我们计算每对等位基因的不同核苷酸所占的比值。例如，等位基因 1－S 和 2－S 在 2379 个核苷酸中相互有 3 个不同，则 $\pi_{12}=0.13\%$。样本中所有等位基因对的 π_{ij} 值列于表 2－1。由等式 2.20，估出核苷酸多样度为 $\pi=0.007$。被研究的等位基因中 6 个是缓慢移动的电泳变异型(S)，5 个为快速移动的电泳变异型(F)。S 和 F 间的区别是由一个氨基酸替换，结果赋于蛋白质以不同的电泳运动性所造成的。对这些电泳类型的每一种，也曾分别地计算了它们的核苷酸多样度. 我们得到：关于 S，$\pi=0.006$；关于 F，$\pi=0.003$。这些结果指出，S 等位基因的多态性是 F 等位基因的两倍。

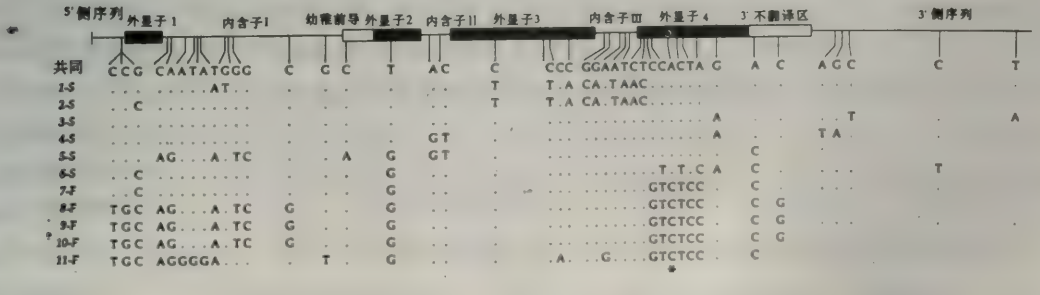


图 2－7　果蝇 *D. melanogaster* 的乙醇脱氢酶基因，其 11 个序列中的多态性核苷酸位点。只有那些与共同顺序有差异的地方被显示出来了。圆点表示与共同顺序相同。外显子 1 中的星号指示赖氨酸-对-苏氨酸替代的位点，该替代是两电泳型等位基因间出现快/慢运动性的原因。自 Hartl 和 Clark(1989) 修改而成。

表 2—1　果蝇 *D. melanogaster*[a] 的醇脱氢酶基因座位的 11 个等位基因中，成对地比较的百分比核苷酸差异

等位基因	等位基因									
	1-S	2-S	3-S	4-S	5-S	6-S	7-F	8-F	9-F	10-F
1-S										
2-S	0.13									
3-S	0.59	0.55								
4-S	0.67	0.63	0.25							
5-S	0.80	0.84	0.55	0.46						
6-S	0.80	0.67	0.38	0.46	0.59					
7-F	0.84	0.71	0.50	0.59	0.63	0.21				
8-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38			
9-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00		
10-F	1.13	1.10	0.88	0.97	0.59	0.59	0.38	0.00	0.00	
11-F	1.22	1.18	0.97	1.05	0.84	0.67	0.46	0.42	0.42	0.42

自 Nei(1987)。数据自 Kreitman(1983)。

a. 加以比较的位点的总数为 2379。S 和 F 分别表示缓慢和快速移动的电泳型等位基因

2.7　新达尔文学说与中性突变假说

达尔文提出他的进化学说根据的是自然选择，而没有关于群体中变异的来源的知识。孟德尔定律被重新发现以及遗传变异被证明是由突变产生的之后，达尔文主义和孟德尔主义被用作后来被称为进化的综合学说(synthetic theory)，或新达尔文主义(neo－Darwinism)的那种理论的框架。根据该学说，虽然突变被看成是遗传变异的根本源，但自然选择却在决定群体的遗传构成，并在基因替换的过程中，被认为起着决定性的作用。

一时，新达尔文主义成了进化生物学中的法则，而选择终于被看成是能驱动进化过程的唯一力量。其他因素，象突变和随机漂变，被认为最多只有次要作用。新达尔文主义的这一特殊标记称为选择主义(selectionism)。

根据选择论者或进化过程的新达尔文学派的观点，基因替换是以选择对有利突变作用的结果而出现的。另一方面多态性则是由平衡选择来维持的。于是，新达尔文论者把替换和多态性看成是由不同的进化力量驱动的、两种孤立的现象。基因替换是正的适应性过程的最后结果，如果且只有当一个新的等位基因能改善该生物的适合度时，它才会经此过程而占据群体的以后世代；而多态性则是在某一基因座位上两个或更多个等位基因共存对该生物或该群体有利时，才被维持的。新达尔文学派的理论主张自然界中大多数遗传多态性是稳定的。

十九世纪七十年代后期群体遗传学中出现了一次革命。蛋白质顺序数据的应用打破了群体遗体学研究中的物种界线，并且首次为检验与基因替换的过程有关的学说提供了足够的经验数据。1968年，木村提出，进化中大多数的分子变化是由于中性或近中性突变的随机固定所造成的(Kimura，1968a 又见；King 和 Jukes，1969)。这一假说，现在以分子进化的中性学说(neutral theory of molecular evolution)而著称，极力主张：分子水平上的大多数进化变化以及物种内的大多数变异性，既不是由有利等位基因的正选择也不是由平衡选择所造成的，而是由选择上呈中性或近中性的突变型等位基因的随机遗传漂变所造成的。从理论意义上讲，中性并不意味着所有等位基因的适合度完全相等。它的意思只是说，这些等位基因的命运很大程度上是由随机遗传漂变所决定的。换句话说，选择可能会起作用，但它的强度太弱以至于不能抵消机遇作用的影响。要使这种情况成为事实，一个等位基因的选择优势度或劣势度的绝对值必须小于 $1/(2Ne)$，其中 Ne 为有效群体大小。

照中性学说的说法，等位基因的频率纯粹由随机规则所决定，而且我们在任一给定时刻得到的画面都只是一种瞬态，它代表取自进行着的动力学过程中的一种暂时构造。因此，多态性基因座位是由或者在走向固定的途中、或者将要灭绝的那些等位基因所组成。从这一场景来看，与进化过程有关的所有分子的表现形式，都应被看成是突变输入和与之相伴的等位基因的随机灭绝或固定的某一连续

过程的结果。故而，中性学说把替换和多态性看成是同一现象的两个侧面。替换是一个长期而渐进的过程，藉此突变型等位基因的频率随机地增加或减少，直到这些等位基因最终因机遇而固定或丢失。在任何给定时间里，某些基因座位所具有的等位基因，其频率将既不是0%也不是100%。这些就是多态性的基因座位。中性学说认为，群体中的大多数遗传多态性在自然界中都是瞬时的。

中性论者和选择论者间争论的本质，涉及突变型等位基因的适合度值的分布问题。两种学说都认为，大多数新突变是有害的并且它们很快被从群体中清除，所以，它们对替换速率和群体内的多态性的量都没有什么贡献。不同的是，关于非有害突变中中性突变的相对比例问题。选择论者主张很少的突变是选择中性的，中性论者却认为大多数非有害突变都是有效中性的。

前20年时间里关于中性突变假说的激烈争论给分子进化带来了很大影响。首先，它导致了在考虑分子变化的进化动力学时，随机漂变的作用不容忽视，这一点得到普遍承认。第二，分子生物学和群体遗传学间综合，通过分子进化和遗传多态性只是同一现象的两个侧面这一概念的导入，而大大加强(Kimurra 和 Ohta，1971)。虽然争论仍在继续，但任何令人满意的进化学说必须与分子水平上进化过程的这两个方面一致，这一点现在已得到承认。

在一系列研究中，根井等(Nei 等，1978)曾从这一观点检验过中性突变假说。最近，赫德森等(Hudson 等，1987)曾提出过一种方法，测试由种间 DNA 顺序比较揭示的高速进化着的 DNA 区域，是否象中性突变假说所预测的那样，在物种内也表现出高水平的多态性。

习题

1. 从等式 2.2 导出等式 2.3。

2. 如果 A_2 对 A_1 是完全显性，那么等位基因 A_2 的每世代频率改变将是多少？

3. 从等式 2.4 导出等式 2.5 中的平衡频率。

4. 给定一个由 5 个两倍体个体组成的群体，其中 A_1 的频率为 0.5，且 A_1 和 A_2 有同样的适合度，那么，在下一世代 A_1 的频率为(a)0.0，(b)0.5，或(c)1.0 的概率各为多少？

5. 在一个雌体以 2∶1 超过雄性的群体中，有效群体大小与调查统计的群体大小的比是多少？

6. 一个经历过瓶颈的群体，比如在连续 6 个世代中其群体大小为：10^4、10^4、10^4、10、10^4 和 10^4。其长期群体大小是多少？

7. 在一个有效群体大小为 100，且 Ne=N 的群体中，一个具 0.01 的选择劣势度的新突变，其固定概率是多少？

8. 应用图 2—7 中的序列，计算等位基因(a)1—S、2—S 或 3—S 间，(b)1—S 和 7—F 间，编码区(黑长方块)中的核苷酸多样度。该编码区长 771 个核苷酸。

后继阅读文献

Christiansen. F. B. and M. W. Feldman. 1986. *Population Genetics*. Blackwell Scientific Publications. Cambridge. MA

Crow. J. F. and M. Kimura. 1970. *An Introduction to Population Genetics Theory*. Harper & Row. New York.

Hartl. D. L. and A. G. Clark. 1989. *Principles of Population Genetics*. 2nd Ed. Sinauer Associates. Sunderland. MA.

Hedrick. P. W. 1983. *Genetics of Populations*. Science Books International. Portola Valley. CA.

Kimura. M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press. Cambridge.

Nei. M. and D. Graur. 1984. Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. 17: 73—118.

3 核苷酸顺序中的进化变化

DNA 序列进化中的基本形式是核苷酸随时间的改变。这一过程值得详细考虑，因为核苷酸顺序中的变化在分子进化研究中，既用来估计进化的速率又用于重建生物进化的历史。然而，核苷酸替换的过程通常是极其缓慢的，以至它不可能在研究者所生存的时间里被观察到。因此，为了检出 DNA 序列中的进化变化，我们依靠比较法，即让某一给定顺序与另一个与它在进化上过去有共同祖先的顺序比较。这种比较要用到统计学方法，其中几种将在本章中讨论。

3.1 DNA 序列中的核苷酸替换

前一章中，我们把进化过程描绘成一系列的基因替换，在这一过程中新等位基因以单个突变的形式产生，继而增加其频率，最终则在群体中固定。现在我们从不同的观点来看这一过程。我们注意到，要被固定的等位基因其顺序不同于它们所替代的等位基因。如果我们使用的时间尺度中一个时间单位比固定时间长，那么，任何给定基因座位上的 DNA 顺序都将表现出连续变化。为此，研究一下一个 DNA 序列中的核苷酸如何随时间改变将是有趣的。象以后我们要解释的那样，这一研究结果可被用来建立估计两序列间替换数的方法。

为了研究核苷酸替换的动力学，我们必须作出几项关于一个核苷酸被另一个替换的概率的假定。文献中已提出了许多这样的数学方案。我们将把讨论仅限制在那些最简单且最常用的方法上：朱克斯和坎托(Jukes 和 Cantor，1969)的一参数模型(one-parameter model)和木村(Kimura ，1980)的两参数模型(two-parameter model)。关于更一般的模型的评论，读者可参考李等(Li 等，1985a)的论述。

朱克斯和坎托的一参数模型

朱克斯和坎托模型的替换方案如图 3－1 所示。该模型假定替换在 4 种核苷酸类型中随机地发生。换言之，在变化方向上没有任何倾斜。例如，如果所考虑的核苷酸是 A，则它将以相同的概率改变

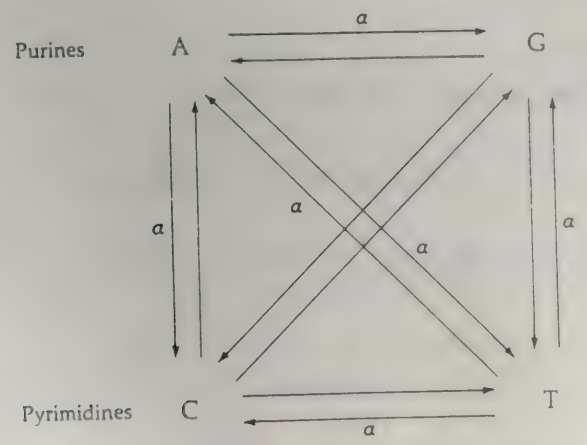


图 3－1 核苷酸替换的一参数模型。在此模型中，每一方向的替换速率都是 α。

成 T、C 或 G。在此模型中，对每种核苷酸来说，替换速率为每单位时间 3α，且 3 种可能的变化方向中每种的替换速率都是 α。因为该模型只涉及一个参数 α，所以它又叫一参数模型。

我们假定一个DNA序列中某一位点上座落的核苷酸在时刻0时为A。首先，我们要问："该位点在时刻t时被A占据的概率是多少？"该概率用$P_{A(t)}$表示。

因为我们从A开始，所以该位点在0时刻被A占据的概率是$P_{A(0)}=1$。在时刻1，该位点上仍为A的概率由

$$P_{A(1)} = 1 - 3\alpha \tag{3.1}$$

给出，它反映出核苷酸保持不变的概率，即$1-3\alpha$。

在时刻2仍有A的概率为

$$P_{A(2)} = (1 - 3\alpha)P_{A(1)} + \alpha[1 - P_{A(1)}] \tag{3.2}$$

为了得出此式，我们考虑两种可能的局面：(1)核苷酸保持不变，和(2)核苷酸已变成T、C或G，但随后又回复到A(图3—2)。在时刻1核苷酸为A的概率是$P_{A(1)}$，而在时刻2保持为A的概率是$1-3\alpha$。这两个独立变量的乘积给出了第一种局面的概率，它构成等式3.2的第一项。在时刻1核苷酸不是A的概率为$1-P_{A(1)}$，而在时刻2变成A的概率为α。这两个概率的乘积给出了第二种局面的概率，它即等式3.2中的第2项。

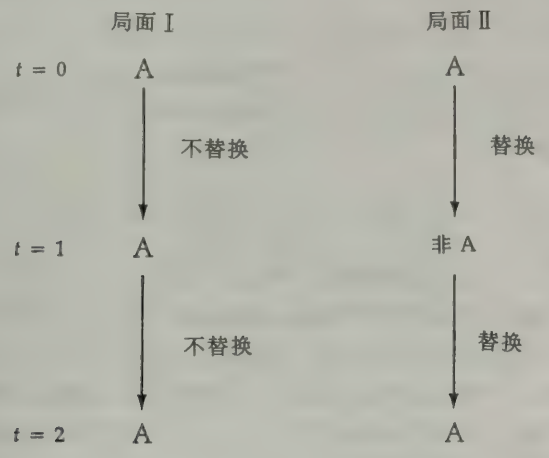


图3—2 假定在时刻0某一位点上为A，而在时刻2该位点上仍有A的两种可能的局面。

用以上公式，我们可以证明，以下递推式可用于任何的t：

$$P_{A(t+1)} = (1 - 3\alpha)P_{A(t)} + \alpha[1 - P_{A(t)}] \tag{3.3}$$

我们可以按每单位时间$P_{A(t)}$的改变量重写等式3.3，为：

$$P_{A(t+1)} - P_{A(t)} = -3\alpha P_{A(t)} + \alpha[1 - P_{A(t)}] \tag{3.4a}$$

或

$$\Delta P_{A(t)} = -3\alpha P_{A(t)} + \alpha[1 - P_{A(t)}] = -4\alpha P_{A(t)} + \alpha \tag{3.4b}$$

至此我们考虑的是一个离散的时间过程。不过，我们可以用连续时间模型来近似这一过程，把$\Delta P_{A(t)}$看成是时刻t时的变化率。以这一近似，等式3.4b被重写成

$$\frac{dP_{A(t)}}{dt} = -4\alpha P_{A(t)} + \alpha \tag{3.5}$$

这是一个一阶线性微分方程，其解由

$$P_{A(t)} = \frac{1}{4} + (P_{A(0)} - \frac{1}{4})e^{-4\alpha t} \tag{3.6}$$

给出。因为我们从A开始，所以$P_{A(0)}=1$。故而

$$P_{A(t)} = \frac{1}{4} + (\frac{3}{4})e^{-4\alpha t} \tag{3.7}$$

事实上，等式3.6不管在什么起始条件下都成立。例如，若该起始核苷酸不是A，则$P_{A(0)}=0$，而在时刻t该位置上有A的概率为

$$P_{A(t)} = \frac{1}{4} - (\frac{1}{4})e^{-4\alpha t} \tag{3.8}$$

等式 3.7 和 3.8 对描绘替换过程来说是充分的。从等式 3.7，我们可以看到，如果起始核苷酸是 A，那么 $P_{A(t)}$ 将呈指数地从 1 降到 1/4（图 3－3）。另一方面，从等式 3.8 我们看到，如果起始核苷酸不是 A，那么 $P_{A(t)}$ 将从 0 单调地上升到 1/4。所以，不管起始条件如何，$P_{A(t)}$ 最终都将达到 1/4（图 3－3）。这对 T、C 和 G 而言也是正确的。因此，在朱克斯-坎托模型下 4 种核苷酸中每种的平衡频率都是 1/4。达到平衡后，在概率上将没有进一步的变化，即，对所有 t 都有 $P_{A(t)}=P_{T(t)}=P_{C(T)}=P_{G(t)}=1/4$。然而，核苷酸的频率仅在无限长的 DNA 序列中保持不变。实际上，DNA 序列的长度是有限的，所以核苷酸频率上的波动看来是会发生的。

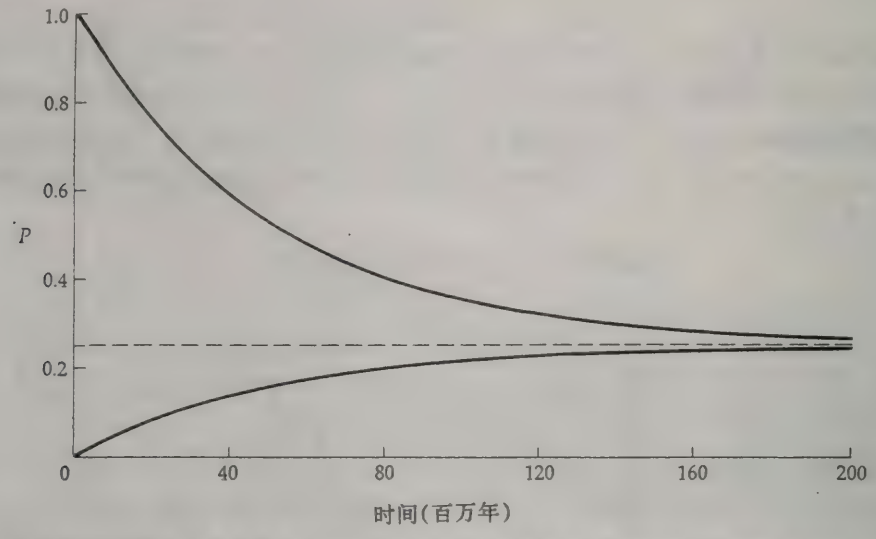


图 3－3　一个位置上有某一核苷酸的概率随时间的变化：由同样的核苷酸开始（上线）或由不同核苷酸开始（下线）。虚线表示平衡频率(0.25)。$\alpha=5\times10^{-9}$ 核苷酸/位点/年。

上面，我们的注意力集中在一个特定的核苷酸位点上，而把 $P_{A(t)}$ 处理为一种概率。然而，$P_{A(t)}$ 也可被解释成某一 DNA 序列中 A 的频率。例如，如果我们从一个仅由腺嘌呤构成的序列开始，那么 $P_{A(0)}=1$，而 $P_{A(t)}$ 则是在时刻 t 该序列中 A 的期望频率。

把起始核苷酸是 A 且时刻 t 的核苷酸还是 A 这一事实考虑进去，我们可以把等式 3.7 以更明确的形式重写成：

$$P_{AA(t)}=\frac{1}{4}+\left(\frac{3}{4}\right)e^{-4\alpha t} \tag{3.9}$$

如果起始核苷酸是 G 而不是 A，那么由等式 3.8 我们得

$$P_{GA(t)}=\frac{1}{4}-\left(\frac{1}{4}\right)e^{-4\alpha t} \tag{3.10}$$

因为在朱克斯-坎托模型下所有核苷酸都是等价的，所以 $P_{GA(t)}=P_{CA(t)}=P_{TA(t)}$。事实上，我们可以考虑一个一般性的概率，$P_{ij(t)}$，这是某一核苷酸在给定起始核苷酸为 i 的条件下在时刻 t 变为 j 的概率。应用这个一般化了的概念和等式 3.9，我们得

$$P_{ii(t)}=\frac{1}{4}+\left(\frac{3}{4}\right)e^{-4\alpha t} \tag{3.11}$$

且由等式 3.10：

$$P_{ij(t)}=\frac{1}{4}-\left(\frac{1}{4}\right)e^{-4\alpha t} \tag{3.12}$$

这里 $i\neq j$。

木村的两参数模型

象朱克斯和坎托模型那样，假定所有核苷酸替换随机发生，这是不现实的。例如，转换（即 A 和 G 之间或 C 和 T 之间的变化）一般比颠换（即所有其他类型的变化）更频繁一些（第 4 章）。考虑到这一事实，木村（Kimura，1980）曾提出一个两参数模型，如图 3－4 所示。在此方案中，每一核苷酸位点上

转换型替换的速率为每单位时间 α，而每种颠换型替换类型的速率则为每单位时间 β。

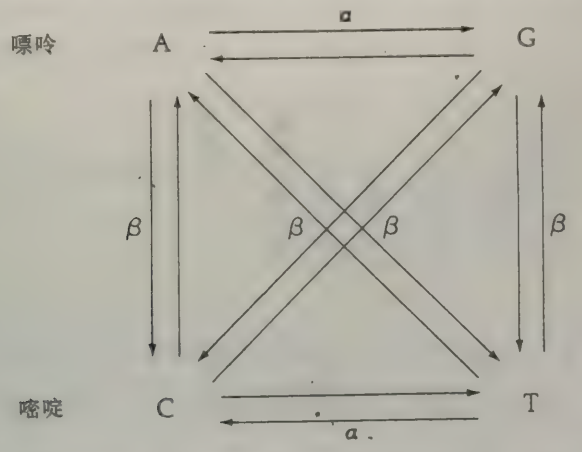


图 3—4 核苷酸替换的两参数模型。在此模型中，转换的速率(α)可能不等于颠换的速率(β)。

该模型比朱克斯-坎托模型复杂，而我们将只给出最后结果。从等式 3.11 我们看到，在朱克斯-坎托模型中，某一位点上在时刻 t 时的核苷酸与时刻 0 时的相同的概率，对 4 种核苷酸来说是相同的。即，$P_{AA(t)}=P_{GG(t)}=P_{CC(t)}=P_{TT(t)}$。由于替换方案的对称，这种等同性对木村的两参数模型也是成立的。我们将用 $X_{(t)}$ 表示该概率。可以证明：

$$X(t)=\frac{1}{4}+(\frac{1}{4})e^{-4\beta t}+(\frac{1}{2})e^{-2(\alpha+\beta)t} \tag{3.13}$$

在朱克斯-坎托模型下，等式 3.12 不管从核苷酸 i 到核苷酸 j 的变化是转换还是颠换都成立。与之不同，在木村的两参数模型下，我们必须对转换和颠换两种变化加以区别。我们用 $Y_{(t)}$ 表示起始核苷酸和时刻 t 时的核苷酸经转换而互不相同的概率。我们看到，由于替换方案的对称，所以 $Y_{(t)}=P_{AG(t)}=P_{GA(t)}=P_{TC(t)}=P_{CT(t)}$ 可以证明

$$Y(t)=\frac{1}{4}+(\frac{1}{4})e^{-4\beta t}-(\frac{1}{2})e^{-2(\alpha+\beta)t} \tag{3.14}$$

时刻 t 时的核苷酸与起始核苷酸经某一特定类型的颠换而互不相同的概率，$Z_{(t)}$，由下式给出：

$$Z(t)=\frac{1}{4}-(\frac{1}{4})e^{-4\beta t} \tag{3.15}$$

注意，每种核苷酸只有一种转换类型，但却经受着两种类型的颠换。例如，若起始核苷酸是 A，那么这两种可能的颠换变化即为 A→C 和 A→T。因此，起始核苷酸与时刻 t 时的核苷酸经两种颠换类型之一的变化而互不相同的概率，将是由等式 3.15 给出的概率的两倍。还要注意，$X_{(t)}+Y_{(t)}+2Z_{(t)}=1$。

3.2 两 DNA 序列间的核苷酸替换数

一个群体中的等位基因替换一般要花成千甚至上百万年来完成(第二章)。为此，我们不能靠直接观察来处理核苷酸替换的过程，核苷酸替换常常是从那些有共同起源的 DNA 分子的成对比较中推断出来的。

两个核苷酸序列相互分歧以后，每一个都将积累核苷酸替换。所以，自两序列发生分歧以来所出现的核苷酸替换数，就是分子进化中最通常用到的变量。

当两核苷酸序列间的分歧程度较小时，在任一位点上发生一次以上替换的机会可以忽略，则两序列间被观察到的差异数将接近实际替换数。另一方面如果分歧程度突出，那么，由于在同一位点上的多重替换(multiple substitution)或多次"击中"(multiple "hits")，观察差异数看来将小于实际替换数。例如，如果某一位点上的核苷酸，在一个序列中从 A 变到 C 再变到 T，在另一个序列中则从 A 变到 T，那么，尽管已发生了 3 次替换，但两序列在该位点上却是相同的(图 3—5)。文献中已有几种修正这

种偏差的方法被提了出来。

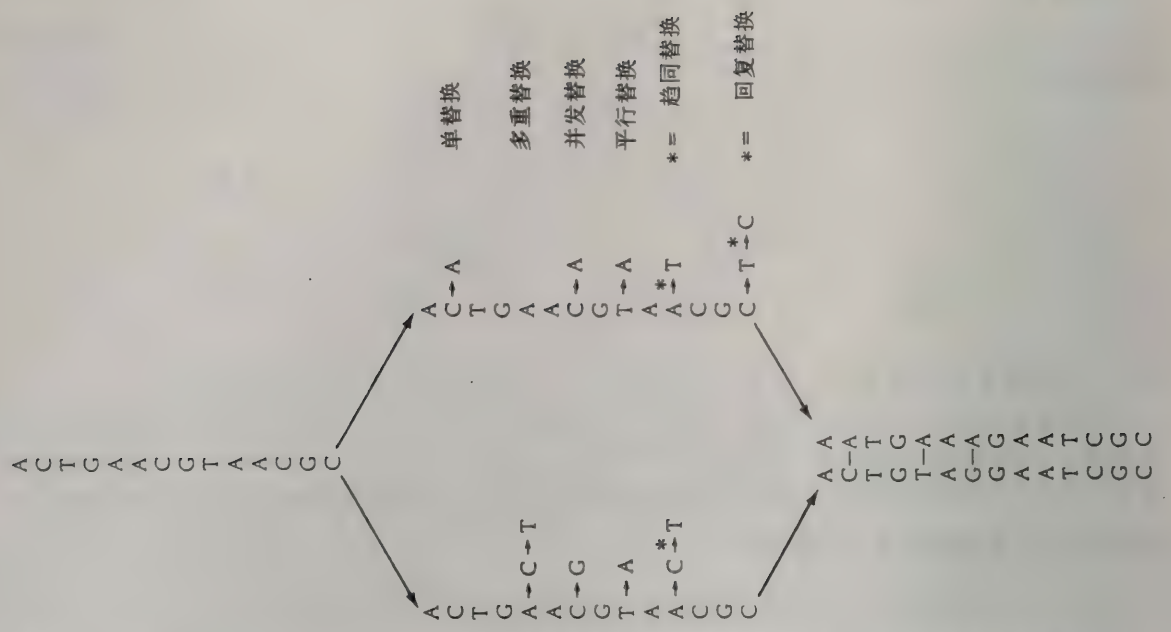


图 3—5 来自一个祖先序列且从它们开始分歧时起累积突变的两个同源 DNA 序列。注意，虽然已经累积了 12 次突变，但可被检出的差异却只有 3 个核苷酸位点。再注意，“并发替换”，“平行替换”，“趋同替换”和“回复替换”，都涉及同一位点上的多重替换，虽然这些替换可能发生在不同的品系中。

替换数通常是以每核苷酸位点的替换数的形式表示，而不是以两序列间的总替换数表示。这有利于长度不相同的序列对间的分歧程度的比较。

为蛋白质编码的序列和非编码序列应分别处理，因为它们通常以不同的速率进化。在前一种情况下，建议将同义替换和非同义替换加以区别，因为已知它们是以显然不同的速率进化着的(第四章)。另一方面，在非编码区，则可假定所有位点以同样的速率进化。

两非编码序列间的替换数

我们在本章前部分对单个 DNA 序列得到的结果，可用于研究两个有共同起源的序列间的核苷酸分歧。我们先从一参数模型开始。在该模型中，只考虑 $I_{(t)}$ 就够了。$I_{(t)}$ 是在时刻 t 某一给定位点上的核苷酸在两个序列中相同的概率。假定某一给定位点上的核苷酸在时刻 0 是 A。在时刻 t，一个后代序列在该位点上有 A 的概率为 $P_{AA(t)}$，因此两个后代序列在该位点上都有 A 的概率则为 $P_{AA(t)^2}$。类似地，两个序列在该位点上都有 T、C 或 G 的概率分别应为 $P_{AT(t)^2}$，$P_{AC(t)^2}$ 和 $P_{AG(t)^2}$ 因此，

$$I_{(t)} = P_{AA(t)^2} + P_{AT(t)^2} + P_{AC(t)^2} + P_{AG(t)^2} \tag{3.16}$$

由等式 3.11 和 3.12，我们得

$$I_{(t)} = \frac{1}{4} + (\frac{3}{4})e^{-8\alpha t} \tag{3.17}$$

等式 3.17 对 T、C 或 G 也成立。因此，不管某一位点上的起始核苷酸如何，$I_{(t)}$ 表示从 t 时间单位以前开始分歧的两个序列间相同核苷酸的比例。注意，在时刻 t 两序列在某一位点上不同的概率为 $P=1-I_{(t)}$。所以，

$$P = \frac{3}{4}(1 - e^{-8\alpha t}) \tag{3.18a}$$

或

$$8\alpha t = -\ln(1 - \frac{4}{3}P) \tag{3.18b}$$

两序列发生分岐的时间通常是未知的，这样我们就不能估出 α。所以我们不去求它而去算 K，K 是两序列间自分岐以来的每位点替换数。在一参数模型的情况下，K＝2(3αt)，这里 3αt 是两个品系的每一个中每位点的替换数。应用等式 3.18b，我们可算出 K 为

$$K = -\frac{3}{4}ln(1-\frac{4}{3}P) \tag{3.19}$$

其中 P 是两序列间不同核苷酸的比例(Jukes 和 Cantor，1969)。对于长为 L 的序列，取样方差由

$$V(k) = \frac{P(1-P)}{L\left(1-\frac{4}{3}P\right)^2} \tag{3.20}$$

近似给出(Kimura 和 Ohta，1972)。

在两参数模型的情况下，两序列间的差异被分类成转换型和颠换型。设 P 和 Q 分别为两序列间转换型和颠换型差异的比例。那么，两序列间核苷酸替换数 K，由

$$K = \frac{1}{2}ln(a) + \frac{1}{4}ln(b) \tag{3.21}$$

来估出，这里 $a=\frac{1}{1-2p-Q}$，$b=\frac{1}{1-2Q}$。取样本方差则由

$$V(K) = \frac{a^2P + c^2Q - (aP+cQ)^2}{L} \tag{3.22}$$

近似给出，其中 c＝(a＋b)/2，而 L 则为这些序列的长度(Kimura，1980)

让我们来考虑一个假想的数字例，设长为 200 的两个序列有 20 个转换和 4 个颠换差异。则 L＝200，P＝20/200＝0.1，Q＝4/200＝0.02。此例据两参数模型我们有：a＝1/(1－0.2－0.02)＝1.28，b＝1/(1－0.04)＝1.04，和 K＝(1/2)ln(1.28)＋(1/4)ln(1.04)≈0.13。替换总数可由每位点替换数 K，乘以位点数 L 求得。在此例中，从两序列间的 24 个差异，我们得到一个约为 26 次替换的估值。按照一参数模型，p＝24/200＝0.12 和 K≈0.13。于是，用一参数模型，我们达到了与两参数模型情况一样的结果。

上例中两种模型基本上给出了同样的估值，这是因为岐化程序低，以至修正后的岐化度(K＝0.13)只略大于未经修正的值(p＝24/200＝0.12)的缘故。在这样的情况下，我们可用较简短的朱克斯和坎托的模型。

当两序列间的岐化度较大时，由两种模型得出的估值就可能差异显著。例如，两个具 L＝200、相互有 50 个转换和 16 个颠换差异的序列，有 p＝50/200＝0.25，Q＝16/200＝0.08。按两参数模型，我们有 a＝2.38，b＝1.19 及 k≈0.48。而根据一参数模型，p＝66/200＝0.33，和 K≈0.43。可见，按一参数模型 K 的估值小于用两参数模型得到的估值。当两序列间的岐化度较大、且特别地在有预先存在的原因相信转换的速率与颠换速率很不一样的情况下，两参数模型看来比一参数模型更精确。

两个为蛋白质编码序列间的替换数

在研究为蛋白质编码序列中，我们通常将起始和终止密码子排除在外，因为这两个密码子几乎不随时间而变。

为了分别处理同义替换和非同义替换，我们首先要把余下的核苷酸位点按以下方式分类：考虑一个密码子中的某一特别位置。设 i 为该位点上可能的同义变化数. 那么该位点被算作 $\frac{i}{3}$ 同义的和 $\frac{3-i}{3}$ 非同义的。例如，在密码子 TTT(Phe)中，最初两个位置被算作非同义的，因为在这两个位置上不发生同义变化；而第 3 位被 算作三分之一同义的和三分之二非同义的，因为该位置上三种可能的变化中有一种是同义的。另一个例子，密码子 ACT(Thr)有两个非同义位点(前两个位置)和一个同义位点(第 3 位)，因为前两位上所有可能的变化都是非同义的，而第 3 位上所有可能的变化都是同义的。在比较两个序列时，我们先要算出每一序列中同义位点的数目和非同义位点的数目，然后计算这两个序列间的平均值。我们用 Ns 表示同义位点的平均数，用 N_A 表示非同义位点的平均数。

其次，我们把核苷酸差异分成同义的差异和非同义的差异两类。对于只有一个核苷酸差异的两个

密码子，这种差异很容易判断。例如，GTC(Val)和GTT(Val)这两个密码子间的差异是同义的，而GTC(Val)和GCC(Ala)这两个密码子间的差异则是非同义的。对于有不止一个核苷酸差异的两个密码子，我们必须考虑导致该观察到的变化的所有可能的进化途径。例如对AAT(Asn)和ACG(Thr)这两个密码子，即有两种可能的途径：

途径Ⅰ：AAT(Asn)↔ACT(Thr)↔ACG(Thr)

途径Ⅱ：AAT(Asn)↔AAG(Lys)↔ACG(Thr)

途径Ⅰ需要一次同义变化和一次非同义变化，而途径Ⅱ则需要二次非同义变化。已知同义替换远比非同义替换发生得频繁(第四章)，所以我们可以假定途径Ⅰ比途径Ⅱ可能性更大一些。例如，如果我们假定途径Ⅰ的权重为0.7而途径Ⅱ的权重为0.3，那么两密码子间同义的差异数估计为$0.7\times1+0.3\times0=0.7$，而非同义的差异数为$0.7\times1+0.3\times2=1.3$。这里所用的权重是假设的。对所有可能的密码子对的权重作经验性的估计，宫田和安永(Miyata和Yasunaga，1980)曾根据蛋白质顺序数据作出，李等(Li等，1985b)则根据DNA顺序数据而得到。如果我们假定两种途径可能性相同，那么，上例的非同义差异数为$(1+2)/2=1.5$，而同义差异数则为$(1+0)/2=0.5$。可见，加权法和非加权法可能会给出有点不同的结果。实际上，两种方法的估值间的差异一般较小(Nei和Gojobori，1986)，但对于那些高度保守的蛋白质，如组蛋白和肌动蛋白，对编码的基因而言它们可能非常重要(Li等，1985b)。用任何一种方法，我们都能估出两编码序列间的同义差异数(Ms)和非同义差异数(M_A)。

从以上结果我们可以用$p_S=Ms/Ns$算出每同义位点的同义差异数，并用$p_A=M_A/N_A$算出每非同义位点的非同义差异数。这些公式显然没有把同一位点上多次击中的效应考虑进去。我们可用朱克斯和坎托的公式：

$$K_s=-\frac{3}{4}ln\left(1-\frac{4Ms}{3Ns}\right) \tag{3.23}$$

和

$$K_A=-\frac{3}{4}ln\left(1-\frac{4M_A}{3N_A}\right) \tag{3.24}$$

来做这样的修正。

一种可采用的处理编码区的方法是，把核苷酸位点分成非简并的(nondegenerate)，两重简并的(twofold degenerate)和四重简并的(fourfold degenerate)位点(Li等，1985b)。如果一个位点上所有可能的变化都是非简并的，则该位点是非简并的；如果三种可能的变化中一种是同义的，则该位点是两重简并的；如果所有可能的变化都是同义的，则该位点即为四重简并的。例如，密码子TTT(Phe)的前两位是非简并的，而第3位则是两重简并的(见第一章中的表1－1)。相比之下，密码子GTT(Val)的第3位是四重简并的。3个异亮氨酸密(Ile)码子中的第3位被简化处理成两重简并位点，尽管事实上该位置上的简并是三重的。在哺乳动物的线粒体基因中，异亮氨酸只有两个密码子，所以其第3位事实上就是两重简并位点(见第一章中的表1－3)。

将核苷酸位点经上述分类分成各种简并类型(degeneracy classes)之后，我们即可对这3类位点分别计算两编码序列间的替换数。注意，根据定义所有非简并位点上的替换都是非同义的。类似地，所有四重简并位点上的替换都是同义的。在两重简并位点上，转换型变化(C↔T和A↔G)是同义的，而所有其他变化，即颠换型变化，都是非同义的。在哺乳动物线粒体的遗传密码里，此规则一无例外。另一方面，在通用的细胞核遗传密码中，却有两个例外：精氨酸密码子(CGA和AGA，CGG和AGG)的第1位，其上的一种颠换型变化是同义的，而其他类型的颠换和所有转换都是同义的；以及3个异亮氨酸密子(AUU、AUC和AUA)中的最后一位也是如此。

根据两种方法计算替换速率的计算机程序可由作者提供，若需要，请寄一个格式化了的IBM PC一兼容软磁盘来拷贝。

3.3 核苷酸序列和氨基酸序列的线性排比

两个同源序列的比较涉及对缺失和插入位置的鉴别问题，因为两个品系从其共同祖先分岐演化

以来，任何一个中都可能发生这类变化。这一过程称为顺序线性排比(sequence alignment)。两个 DNA 序列的比较通常不能告诉我们，是其中一个序列中发生了丢失呢还是另一个序列中出现了插入。因此，这两类事件的后果统称为裂缝。

虽然我们是用 DNA 序列来说明线性排比的过程，但同样的原则和程序也可用于氨基酸序列的排比。事实上，用氨基酸顺序与用 DNA 顺序比起来，前者通常能得到更可靠的线性排比。

线性排比由一系列成对的碱基组成，其中每一个碱基各来自一个序列。有 3 种线列的对：(1)匹配的碱基对，(2)匹配错误的碱基对，和(3)由来自一个序列的碱基与另一序列的空缺碱基(null base)组成的对子。空缺碱基用—表示。一个匹配的对子表示一个自两序列分岐以来没有发生变化的位点，一个匹配错误的对子表示一次替换，而一个空缺对子则表示，在这两个序列之一的该位置上曾经发生过一次缺失或者插入。

考虑两 DNA 序列 A 和 B，其长度分别为 m 和 n 的例子。如果我们用 x 表示匹配的对子数，用 y 表示匹配错误的对子数，而用 z 表示含有一个空缺碱基的对子数，则我们有：

$$n + m = 2(x + y) + z \tag{3.25}$$

点阵法

当只有少数裂缝且两序列在其他任何方面差异都不太大时，一种合理的线性排比可以由视觉观察得到，或者也可用被称为点阵法(dot matrix method)的方法得到。在此法中，被线排的两个序列作为一个矩阵的首列和首行而写出(图 3—6)。在两序列中核苷酸相同的矩阵位置处记上圆点。如果两序列等同，那么该矩阵对角线上的所有元素都将是圆点(图 3—6a)。如果两序列有差异但可被无裂缝地线排，则对角线元素的大多数是圆点(图 3—6b)。如果两序列之一中出现一个裂缝，则线性排比的对角线将垂直或水平地移动(图 3—6c)。如果两序列间的差异既有裂缝又有替换(图 3—6d)，则找出裂缝的位置并从几种可能的线性排比中挑出一种可能是很困难的。在这样的情况下，视觉观察和点阵法就不可靠了，而为了得到客观的线性排比已有几种计算方法被提出来了

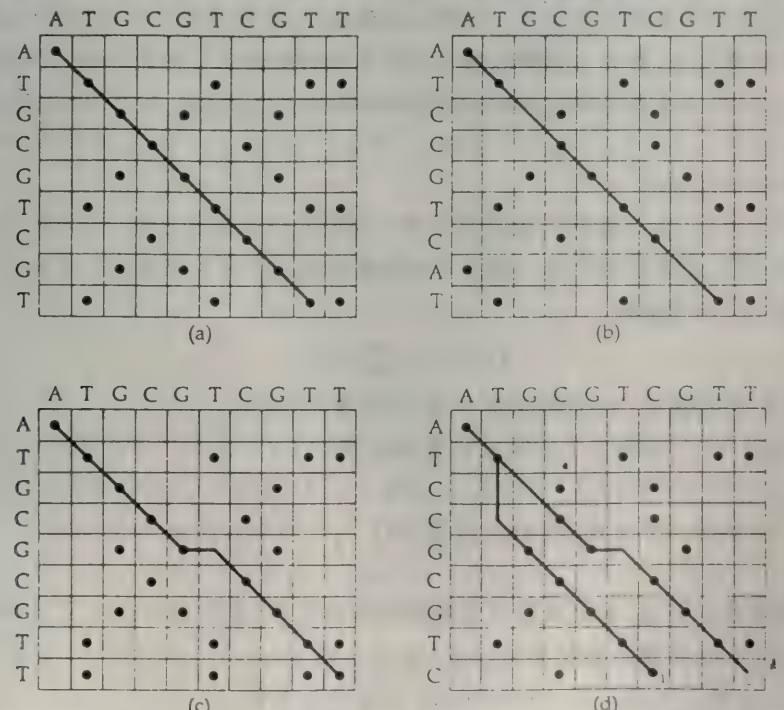


图 3—6　用于线排核苷酸顺序的点阵。(a)两序列等同；(b)两序列有差异但不含裂缝；(c)两序列含有一个裂缝，但此外则相互等同；(d)两序列既有替换又有裂缝。在(d)中，途径 1 由 6 个对角线步骤，其中无空格，和 2 个垂直步骤组成。途径 2 含有 8 个对角线步骤，其中 2 个是空格，和 1 个水平步骤。途径 1 和途径 2 间的选取靠裂缝处罚来决定，即根据哪种进化序列更有可能：一次两核苷酸缺失(途径 1)或一次一核苷酸缺失和两次替换(途径 2)来决定。

顺序—距离法

两序列间最为可能的线性排比是，根据某种标准使线列中匹配错误和裂缝的数目最小的那种。不幸地是，降低匹配错误数结果常会导致裂缝数增加，反之亦然。

例如，考虑以下两个序列：

A：TCAGACGATTG (m=11)

B：TCGGAGCTG (n=9)

我们可按如下排比将匹配错误数降到零：

(Ⅰ) T C A G－A C G－A T T G
T C－G G A－G C－T－G

在这种情况下裂缝数为 6。反之，裂缝数可降低到由|m－n|个核苷酸组成的一个裂缝，结果匹配错误数却增加了：

(Ⅱ) TCAGACGATTG
TCGGAGCTG－－

在这种情况下，我们只有一个位于末端，因而也是不可避免的裂缝，但匹配错误(用星号标出)的数目却为 5。

或者，我们可以选一个裂缝数和替换数都不是最小的线性排比。例如，

(Ⅲ) T C A G－A C G A T T G
T C－G G A－G C T G－

在这种情况下匹配错误数是 2，裂缝数为 4。

那么，这 3 种线性排比中哪一种最可取？显然，将替换与裂缝比较就好象将苹果和桔子比较一样。所以，我们必须找到一个共同标准，藉此来比较裂缝和替换。此共同标准被称为裂缝处罚(gap penalty)。

有几种指定裂缝处罚的系统。所有系统都是在相对于点状的替换、缺失和插入出现的频繁程度如何，这类问题的预先理解的基础上建立的。在第 1 个系统中，裂缝的总长度(z)用恒定的裂缝处罚(w)来乘。该系统背后的假定是，有某一裂缝的概率反比于裂缝的大小。举例说来，有一个由两核苷酸组成的裂缝的概率，与有两个各由一核苷酸构成的裂缝的概率相同。这样，对任何线性排比，我们都能用

$$D = y + wz \tag{3.26}$$

来计算两序列间的距离尺度(D)。

在第 2 种处罚系统中，我们假定长的缺失和插入在进化中与短的比较，出现的可能性是不同的。在这种情况下，对不同长度裂缝的处罚可能正比于裂缝长度也可能并非如此。根据这一系统，与某一特定线性排比有关的距离尺度是

$$D = y + \sum w_k z_k \tag{3.27}$$

其中，zk 是长度为 K 的裂缝数，wk 则是对长为 K 的裂缝的处罚。

现在让我们用有 w=2 的第 1 个系统来比较线性排比Ⅰ，Ⅱ和Ⅲ。得到的距离(D)，对线性排比Ⅰ，Ⅱ和Ⅲ分别为：0+(2×6)=12，5+(2×2)=9，和 2+(2×4)=10。我们将选取线排Ⅱ。如果我们用有 $w_1=2$ 和 $w_2=6$ 的第 2 个系统，则 D 的值结果对Ⅰ、Ⅱ和Ⅲ分别为 12，11 和 10。在这种情况下，我们选取线排Ⅲ。

任何线性排比算法的目的，都是从所有可能的线性排比中，选取具有最小 D 值的那种线性排比。在最常应用的方法中，有尼德尔曼与文施(Needleman 和 Wunsch，1970)法，和塞勒斯(Seller，1974)法．在前一种方法中，两序列间的类似性(similarity)用类似指数(similarity index)来测度，而具最大类似性的那种线性排比将被从所有候选者中选取出来。在塞勒斯法中，两序列间的不相似性(dissimilarity)用距离指数(distance index)来测度，具最小距离的那种线性排比将被选出。这两种方法曾被证明在某些条件下是等价的(Smith 等，1981)。

在必须从许多线性排比中选出一种时，寻找最佳排比的任务若无计算机的帮助常常难以完成。在

尼德尔曼与文施(Needleman 和 Wunsch,1970)算法或其修订法的基础上,已有许多关于线排顺序的常用计算机程序建立起来。

要记住的最重要的一点是,作为最后结果的线性排比常有赖于裂缝处罚的选取,而后者又有赖于,相对于点替换的频率裂缝事件在 DNA 和蛋白质进化中的频率究竟是多少的这样一些关键的假定上。

3.4 核苷酸替换数的间接估计

在估计两序列间核苷酸替换数方面,最完全的解决可通过比较它们的核苷酸顺序而得到。不过,替换数也可从其他类型的分子数据,象限制酶图谱或者 DNA—DNA 杂交得到的数据间接地推断出来。

限制性核酸内切酶片段模式和位点图谱

限制性核酸内切酶(restriction endonucleases)或限制酶(restriction enzymes)能识别被称为识别顺序(recognition sequences)的特殊双链 DNA 序列,并在识别顺序上或其近旁切开该 DNA。识别顺序通常长为 4 或 6 碱基对,它们中许多都是回文(即它们是旋转对称的)。识别顺序可能是唯一的(例如 EcoRI),也可能不是唯一的(例如 Hind Ⅱ)(见表 3—1)。切点称为拼接位点(splicing site)或限制位点(restriction site)。许多限制性内切核酸酶以一种错开的方式切开双链 DNA,所以将产生"粘性末端"(sticky ends),以后它们可在连接酶(ligase)的作用下相互连接(ligated)。这就是为什么限制酶能在遗传工程中成为一种极有用的工具的原因。表 3—1 列出了几种限制酶的识别顺序和拼接位点。

表 3—1 几种限制性核酸内切酶的识别顺序和切点

酶 (生物来源)	识别位点	识别顺序(RS)				切割	
		大小	不确定性	回文	邻接	在 RS 中	错开式
EcoR I (*Escherichia coli*)	5′—G↓A—A—T—T—C—3′ 3′—C—T—T—A—A—G↑—3′	6	—	+	+	+	+
Hind Ⅱ (*Haemophilus influenzae*)	5′—G—T—Py↓Pu—A—C—3′ 3′—C—A—Pu↑—Py—T—G—5′	6	+	+	+	+	—
Hae Ⅲ (*Haemophilus aegyptus*)	5′—G—G↓C—C—3′ 3′—C—C↑—G—G—5′	4	—	+	+	+	—
Bbv I (*Bacillus brevis*)	5′—G—C—A—G—C—(N_8)↓—3′ 3′—C—G—T—C—G—(N_{12})↑—5′	5	—	—	+	—	+
Nci I (*Neisseria cinerea*)	5′—C—C↓C/G—G—G—3′ 3′—G—G—G/C↑—C—C—3′	5	+	+	+	+	+
Not I (*Nocardia otitidis-caviarum*)	5′—G—C↓G—G—C—C—G—C—3′ 3′—C—G—C—C—G—G↑—C—G—5′	8	—	+	+	+	+
Hinf I (*Haemophilus influenzae*)	5′—G↓A—N—T—C—3′ 3′—C—T—N—A↑—G—5′	4	—	+	—	+	+

a、识别顺序用黑体字母表示。切点用箭头指出。不确定的地方,象 Pu:嘌呤;Py:嘧啶;C/G:C 或 G;N:任何核苷酸。Nn 表示由 n 个任意的核苷酸组成的序列。

当一个双链的 DNA 片段受到水解时,即有各种不同长度的片段产生出来。它们可因其各自的长度而在电泳凝胶上分开,因为在凝胶上较短的片段比较长的要跑得更快也移动得更远。通过用已知长度的 DNA 片段作基准,限制性片段的长度即可被估计出来。不同的 DNA 序列根据其识别位点的数目和位置的差异而受到限制酶的不同切割。由一个 DNA 序列水解产生的片段的数目和大小被称为限制片段模式(restriction-fragment pattern)。连续而交互地应用几种能将 DNA 水解成重叠片段的限制酶,常常可推断出 DNA 上限制位点的大概位置(图 3—7)。表示某一 DNA 序列上限制位点的位置

的方案图称限制图谱(restriction map)。

应用限制酶来推断两序列间的替换数,其背后的理由是,两 DNA 序列的类似性越大则其限制片段模式就越相似。通过对 DNA 序列内限制位点的分布作出某些假定,例如,象 4 种核苷酸有相同的频率以及它们在序列中的空间分布是随机的,这样的假定,则从限制位点数据就可以对 DNA 序列间限制模式方面的进化变化进行研究,从而估计出每位点的核苷酸替换数(K)。

首先,我们考虑从限制片段模式来估计 K。从共有片段数来估计 K,要求我们对由限制性核酸内切酶水解的 DNA 的电泳模式进行直接比较。这里提供的方法是由根井和李(Nei 和 Li,1979)创导的。文献中曾报导过另外两种方法(Upholt,1977;Engels,1981a),而卡普兰(Kaplan,1983)曾证明了这 3 种方法给出类似的结果。

DNA 的两序列间共有 DNA 片段的期望比例(F),可由

$$\hat{F} = \frac{2m_{XY}}{m_X + m_Y} \tag{3.28}$$

来估计,其中 m_X 和 m_Y 分别是序列 X 和 Y 水解后产生的限制片段的数目,而 m_{XY} 则是两序列共有的片段数。

根井和李(Nei 和 Li,1979)曾证明,共有片段的期望比例(F)可用在 t 时间内某一限制位点保持不变的概率(G)来表示,两者之间的近似关系为

$$F \approx \frac{G^4}{3 - 2G} \tag{3.29}$$

这里,$G = e^{-r\lambda t}$,G 中的 r 为识别位点中核苷酸的数,λ 是核苷酸的替换速率,t 是两序列间发生分岐的时间。这些序列间每位点的替换数为 $K = 2\lambda t$。为了估出 G,我们重新安排等式 3.29,得

$$G = [F(3 - 2G)]^{\frac{1}{4}} \tag{3.30}$$

该方程可通过一个反复尝试过程解出。根井(Nei,1987) 建议用 $G = F^{1/4}$ 作为最初尝试值。一般只需要很少几次反复循环。G 的估值可使我们得到 K 的估值,关系如下:

$$K = -\frac{2}{r} ln(G) \tag{3.31}$$

让我们来考虑下面这样一个例子:取自两种野生小麦(*Aegilops sharonensis* 和 *Ae. bicornis*)的相应线粒体 DNA 片段,用 3 种限制酶,*Bam* Ⅰ, *Hind* Ⅲ 和 EcoRI 来水解,它们的识别顺序都为 6 碱基对长(数据自 Graur 等,1989a)。*Ae. sharonensis* 水解产生 4 个片段,而 *Ae. bicornis* 水解则产生 5 个片段。两个片段为两种小麦所共有。用等式 3.28,我们估出 F 为 2/9=0.222。现在我们可以开始由等式 3.30 给出的反复尝试过程。我们采用的 G 的初值为 $0.222^{1/4} = 0.687$。第一次循环后我们得 G=0.775,而下一次循环 G=0.753。随着尝试的进行摆幅将越来越小,而在第 5 次和第 6 次循环后我们都得到 G=0.758。因此,我们终止反复尝试过程。为了得到两序列间替换数的估值,我们用等式 3.31。最后结果是,两线粒体序列的相互差异用每核苷酸位点替换数 K=0.092 来估计。

现在我们考虑由限制位点图谱估计两序列间的核苷酸替换数。在前面的例子中,限制位点的位置是未知的。如果限制位点已在 DNA 序列上定位,那么,我们可以直接从图谱上找出共有和非共有的位点,并估出替换数。设 m_X 和 m_Y 分别为 DNA 序列 X 和 Y 中限制位点的数目,而 m_{XY} 为两序列间共有的限制识别位点数。X 和 Y 在某一给定位点上共有同样的识别顺序的概率用 S 表示,此值可用

$$\hat{S} = \frac{2m_{XY}}{m_X + m_Y} \tag{3.32}$$

估出(Nei 和 Li,1979)。核苷酸差异的比例,p,可用

$$\hat{p} = 1 - \hat{S}^{\frac{1}{r}} \tag{3.33}$$

估计,其中 r 是识别顺序中核苷酸的数目。两序列间的每位点替换数可用等式 3.19 从已知的 $\hat{p}$ 估出。

限制位点图谱法比限制片段模式法要乏味一些,但却可靠得多。前者在 K 值高达 0.25 时仍可应用,而后者对于 K>0.05 的情况就可能是不精确的。

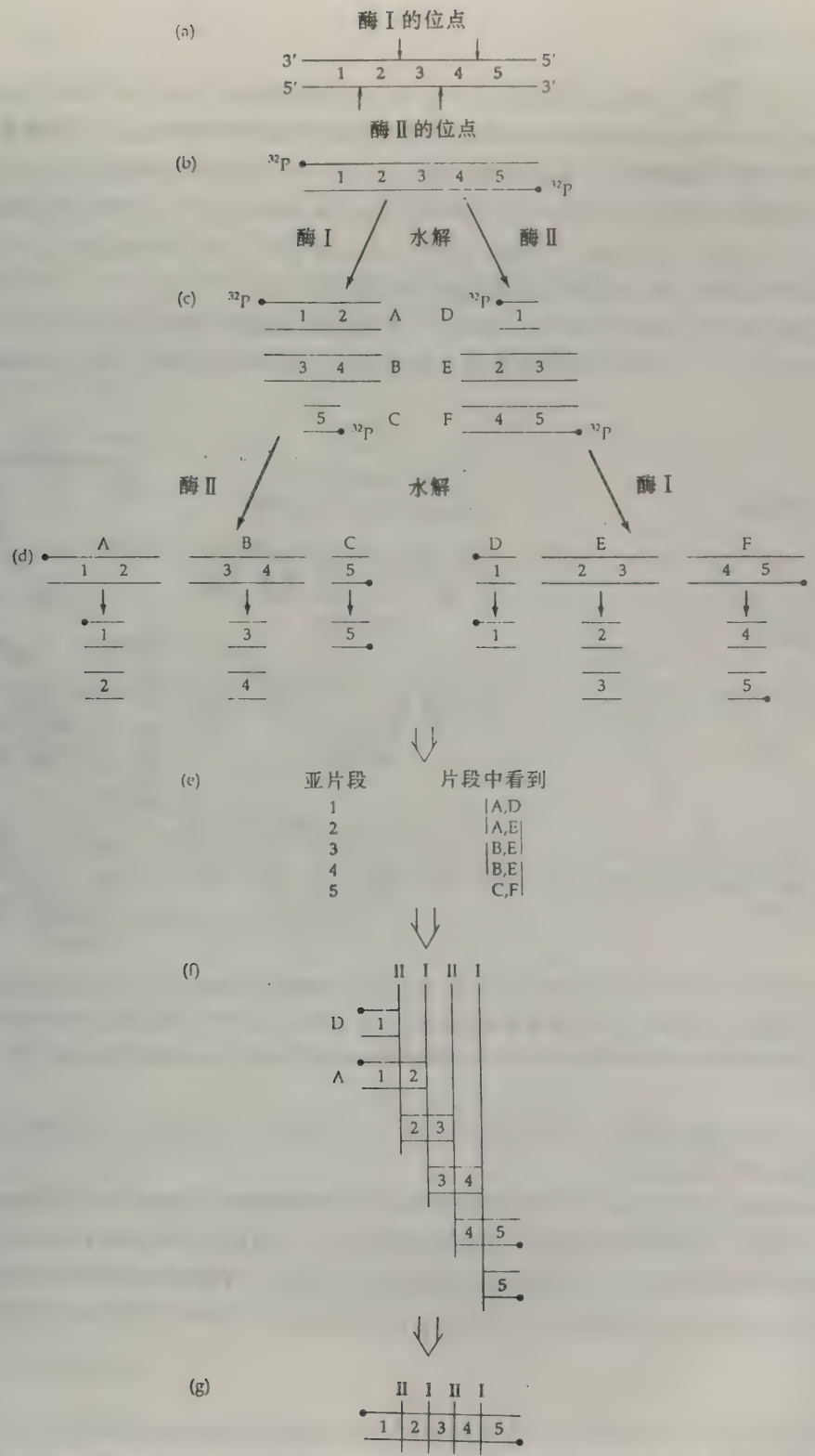


图 3－7　DNA 序列上限制位点的定位图。(a)一个假想的 DNA，具有由两种不同的限制酶识别的识别位点，该序列的识别图谱未知。(b)3′端用放射性标记。(C)DNA 由酶Ⅰ水解产生片段 A、B 和 C，而由酶Ⅱ水解则产生片段 D、E 和 F。(d)用一种限制酶得到的每一片段再用另一种限制酶水解，产生出亚片段 1－5。(e)这些亚片段被用于鉴别重叠片段。(f)这些亚片段按它们间的重叠模式指示的秩序排列。(g)推理出的 DNA 序列的限制图谱。片段和亚片段可根据它们的长度一个个地鉴别出来，而长度则是从它们在电泳凝胶上的位置推出的，末端片段和亚片段则由其放射性标记来鉴别。自 Suzuki 等(1989)修改而成。

DNA－DNA 杂交

DNA－DNA 杂交(DNA－DNA hybridization)技术是以这样的事实为根据的：双链 DNA 分子的热稳定性有赖于两条链间核苷酸匹配的比例。随着匹配比例的降低，双链的热稳定性也降低。在两条链来自同一序列的双链 DNA(即同源双链(homoduplex)分子)中，匹配的比例根据定义应为 100%。另一方面，在两条链来源不同的双链 DNA(即异源双链(heteroduplex)分子)中，匹配的比例则小于 1。其大小有赖于这两个序列自它们从某一共同祖先分化以来究竟累积了多少核苷酸差异。所以，异源双链 DNA 将会在比同源双链 DNA 低的温度下变性或熔解成单链。

DNA 杂交试验的基本实验程序如图 3－8 所示。大致上，在重复序列被除去以后，该程序包括将来自两不同物种的变性 DNA 的混合物缓慢冷却，以制造出人工杂种 DNA 分子。然后，将该混合物逐

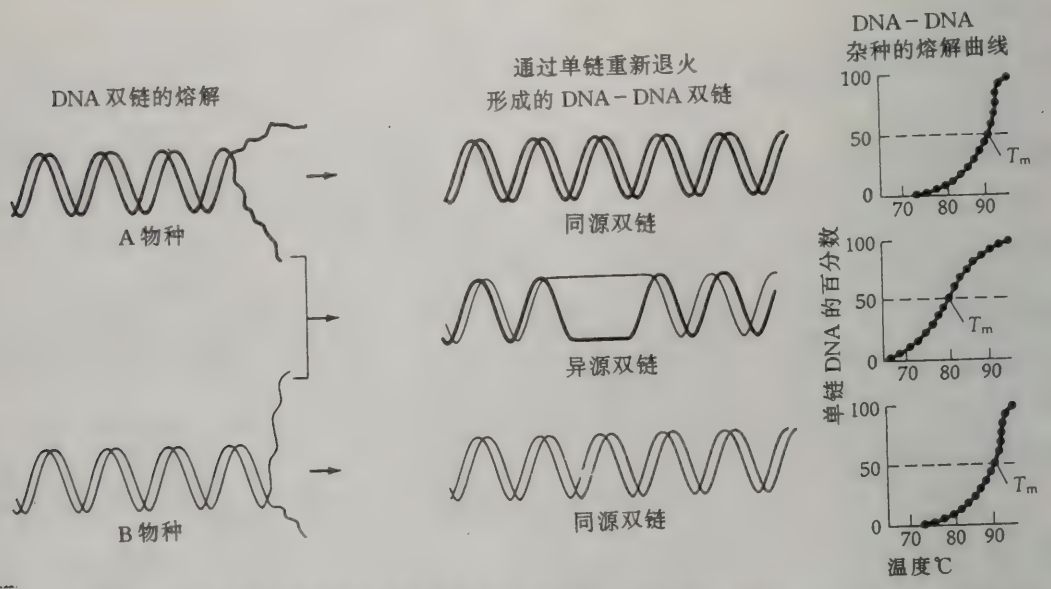


图 3－8　从 DNA－DNA 杂交研究推论出的顺序类似性。双链分子被熔解成单链的 DNA。同源双链和异源双链通过单链的重新退火而形成。50%的 DNA 熔解成单链的温度用 Tm 表示，测定两种同源双链和异源双链的 Tm。两种同源双链间的 Tm 值可能是不同的，同样两互不相同的异源双链类型间的 Tm 值也可能是不同的。自 Avers (1989)修改而成。

渐加热，并在每一温度下测定溶液中单链 DNA 的百分比。关于此法(TEACL 法)的详细介绍，可见例如享特等(Hunt 等，1981)的论述。

杂种 DNA 的热稳定性，用 50%的杂种 DNA 解离成单链时的温度来度量。然后将此半熔解温度与 50%的同源双链 DNA 变成单链时的温度比较。注意，在每一次种间比较中，我们有两种同源双链，每物种各有一种，所以，习惯上我们用它们的半熔解温度的平均值。同源双链和异源双链的半熔解温度间差异，ΔTm，由经验证明，与碱基对误配的比例近似地线性相关(Britten 等，1974)。我们将这种关系表示成

$$p = C\Delta Tm \tag{3.34}$$

这里 p 是误配的比例，C 是一个常数。C 的值通过对碱基对误配度已知的异源双链进行 DNA－DNA 杂交试验，而从经验上得到。C 值被发现随实验条件的不同而在 C＝0.01 和 C＝0.015 之间变化。已知 ΔTm 的实验误差是非常大的，因此，对同样的物种对应该做许多次重复观察。

现在让我们来考虑下面的一个数字例(数据来自 Caccone 和 Powell，1989)。来自人类和矮黑猩猩(*Pan paniscus*)雄性的同源双链 DNA 的平均 Tm 值，分别为 59.50℃和 59.12℃。于是，同源双链分子的 Tm 平均值为 59.31℃。两交互的异源双链 DNA 的 Tm 平均值则为 57.59℃。因此，ΔTm 为 1.72℃。由等式 3.34，我们得到一个每核苷酸位点大约 0.017－0.026 次替换的差异。

习题

1. 证明等式 3.3 对 t=0 成立，即，若 t=0 则它将简化成等式 3.1。

2. 导出等式 3.10，并证明在朱克斯-坎托模型下 $P_{GA(t)}=P_{CA(t)}=P_{TA(t)}$。

3. 当 α=β 时，木村的两参数模型将变得与朱克斯和坎托的一参数模型等同。为了证实这一点。证明：当此条件满足时，等式 3.13 将变得与等式 3.11 相同，且等式 3.14 和 3.15 都变得与等式 3.12 相同。

4. 用等式 3.13，3.14 和 3.15 证明，在木村的两参数模型下一个序列中 4 种核苷酸的平衡频率都是相同的(即 1/4)，这与一参数模型的结果一样。

5. 从等式 3.16 导出等式 3.17。

6. 对以下两个序列：

Ser	Thr	Glu	Met	Cys	Leu	Met	Gly	Gly
TCA	ACT	GAG	ATG	TGT	TTA	ATG	GGG	GGA
TCG	ACA	GGG	ATA	TAT	CTA	ATG	GGT	ATA
Ser	Thr	Gly	Ile	Tyr	Leu	Met	Gly	Ile

计算(a)每同义位点的同义替换数和(b)每非同义位点的非同义替换数。

7. 根据木村的两参数模型，两序列间的差异数为 P+Q，证明当转换和颠换合起来考虑时等式 3.21 将被简化成等式 3.19。

8. 用点阵法线性排比以下两个顺序：

AATGCTTGCATGGGGCTAGTT

ATTGCTGCATGAGGCGCGCTAGT

选出两种可能的线性排比，并决定用每核苷酸为 2 的恒定裂缝处罚时哪一种要更好一些。若用更大的裂缝处罚，比如说 10，该选择会受到影响吗？

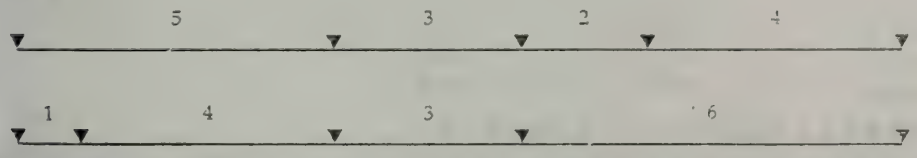


图 3—9　两限制性内切核酸酶图谱的假想例。序列上的数字代表这些片段的长度(以 kb 为单位)。

9. 从图 3—9 中的两序列的限制位点图谱，估计两序列间 核苷酸的替换数，用(a)共有片段的比例，和(b)共有限制位点的比例。限制性内切核酸酶的识别顺序为 4 个核苷酸。用有 6 核苷酸识别位点的限制性内切核酸酶，将会有什么样的结果？该差异的原因是什么？

后继阅读文献

Doolittle. R. F. 1990. *Molecular Evolution: Computer Analysis of Protein and Nucleic Acid Sequences*. Academic Press. San Diego. CA

Li. W.-H.. C. -C. Luo and C. -I. Wu. 1985. Evolution of DNA sequences . pp. 1—94. In R. J. MacIntyre (ed.). *Molecular Evolutionary Genetics*. Plenum. New York.

Nei. M. 1987 *Molecular Evolutionary Genetics*. Columbia University Press. New York.

4 核苷酸替换的速率和模式

前一章中导出的数学理论可用于核苷酸替换速率的研究之中，而该速率则是分子进化研究里的一个基本量。事实上，为了阐明某一DNA序列进化的特性，我们需要知道，它进化得究竟有多快，以及其组成部分的核苷酸替换速率是多少。比较一下基因和不同DNA区域间的替换速率也是很有趣的，因为这可以帮助我们理解进化中核苷酸替换的机制。知道了核苷酸替换的速率，还使我们能对物种间的分岐演化这样的进化事件，给出一个时间年代来。不过，要想做到这一点，我们必须知道从一组物种估出的速率是否能适用于另一组生物种群。这就提出了这样一个问题，即，速率在不同的进化谱系间是怎样变化的。

4.1 核苷酸替换的速率

核苷酸替换的速率(rate of nucleotide substitution)被定义成每年每位点的替换数，并可用两同源序列间的替换数K，除以2T来算出，这里T是两序列间发生分岐的时间(图4—1)。即，

$$r = \frac{K}{2T} \tag{4.1}$$

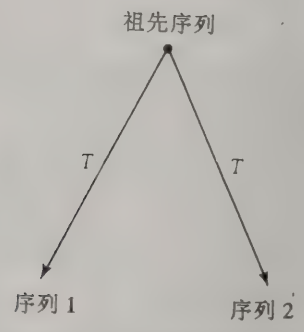


图4—1 两同源序列在T年前从某一共同祖先序列分岐而来。

两序列发生分岐的时间T，假定与两物种发生分化的时间相同，且通常都用古生物学数据资料来推算。

本节我们将处理不同基因间，和某一基因的不同区域间的速率变异问题。为此目的，建议对所有被考虑的基因使用同样的物种对。这有两重原因。首先，关于分岐时间的古生物学估计通常都有很大的不确定性。用同一对物种，我们就可以无须知道分岐时间而去比较各基因间的进化速率。其次，替换速率在各谱系间可能变化很大(见第48页)，在这种情况下，两基因间速率上的差异可能是由谱系间的差异所造成，而不是由两基因本身的差异所造成的。

目前研究核苷酸替换速率的最合适的数据来自哺乳动物，这是因为，有关哺乳动物的DNA序列的数据最为丰富，有关哺乳类的化石记录相对而言特征比较明确、全面，再加上可以得到哺乳类之间相当可靠的分岐时间的缘故。

编码区

我们在表4—1中列出了36种为蛋白质编码的基因的同义替换速率和非同义替换速率。这些速率是从人类与啮齿类同源基因间的比较中算出的。根据与真兽亚纲哺乳类的辐射演化有关的古生物学证据，人类-啮齿类的分岐时间已被设定为8000万年前。

表 4—1　哺乳类各为蛋白质编码的基因的同义替换速率和非同义替换速率[a]

基因	L[b]	非同义替换速率($\times 10^9$)	同义替换速率($\times 10^9$)
组蛋白			
组蛋白 3	135	0.00±0.00	6.38±1.19
组蛋白 4	101	0.00±0.00	6.12±1.32
收编系统蛋白			
肌动蛋白 α	376	0.01±0.01	3.68±0.43
肌动蛋白 β	349	0.03±0.02	3.13±0.39
激素、神经肽和其他活性肽			
生长激素释放抑制因子—28	28	0.00±0.00	3.97±2.66
胰岛素	51	0.13±0.13	4.02±2.29
促甲状腺素	118	0.33±0.08	4.66±1.12
胰岛素样生长因子 Ⅰ	179	0.52±0.09	2.32±0.40
促红细胞生成素	191	0.72±0.11	4.34±0.65
胰岛素 C 肽	35	0.91±0.30	6.77±3.49
甲状旁腺素	90	0.94±0.18	4.18±0.98
促黄体生成激素	141	1.02±0.16	3.29±0.60
生长激素	189	1.23±0.15	4.95±0.77
尿激酶—血纤蛋白溶酶原活化因子	435	1.28±0.10	3.92±0.44
白细胞中介素 Ⅰ	265	1.42±0.14	4.60±0.65
松弛肽	54	2.51±0.37	7.49±6.10
血红蛋白和肌红蛋白			
α-珠蛋白	141	0.55±0.11	5.14±0.90
肌红蛋白	153	0.56±0.10	4.44±0.82
β-珠蛋白	144	0.80±0.13	3.05±0.56
载脂蛋白			
E	283	0.98±0.10	4.04±0.53
A—Ⅰ	243	1.57±0.16	4.47±0.66
A—Ⅳ	371	1.58±0.12	4.15±0.47
免疫球蛋白			
IgV_H	100	1.07±0.19	5.66±1.36
$Ig\gamma_1$	321	1.46±0.13	5.11±0.64
Igk	106	1.87±0.26	5.90±1.27
干扰素			
α1	166	1.41±0.13	3.53±0.61
β1	159	2.21±0.24	5.88±1.08
γ	136	2.79±0.31	8.59±2.56
其他蛋白质			
醛缩酶 A	363	0.07±0.03	3.59±0.52
羟黄嘌呤磷酸核糖基	217	0.13±0.04	2.13±0.35
转移酶			
肌酸激酶 M	380	0.15±0.03	3.08±0.37
甘油醛-3-磷酸脱氢酶	331	0.20±0.05	2.84±0.37
乳酸脱氢酶 A	331	0.20±0.04	5.03±0.61
乙酰胆碱受体 γ 亚基	540	0.29±0.04	3.23±0.31
血纤蛋白原 γ	411	0.55±0.06	5.82±0.67
白蛋白	590	0.91±0.07	6.63±0.61
平均[c]		0.85(0.73)	4.61(1.44)

a. 所有速率都以人和啮齿类基因间的比较为根据，且分歧时间设定为 8000 万年以前。速率以每 10^9 年每位点替换数为单位。

b. L＝受比较的密码子数。

c. 平均指算术平均值，括号内的值为标准偏差，都是用所有基因的值算出的。

我们注意到，非同义替换的速率在基因间变化极大。变化幅度从组蛋白 3 和组蛋白 4 的有效数字为零，到干扰素 γ 的每年每非同义位点 2.79×10^{-9} 替换。某些激素（例如生长激素释放抑制因子—28和胰岛素）是极其保守的，而另一些激素则或者以中间速率（例如促红细胞生成素）进化，或者以高速率（例如白细胞中介素 Ⅰ 和松弛肽）进化。血红蛋白和肌红蛋白以中速进化，而载脂蛋白和免疫球蛋白则进化得非常迅速。

同义替换的速率变化也相当大，不过比起非同义替换来速率变化要小得多。可以证明，基因间同义替换速率方面的变异明显地大于仅由统计学波动造成的期望变异。

对表 4－1 中绝大多数基因来说，同义替换的速率大大超过非同义替换的速率。如在一个最极端的例子，组蛋白 3 中，虽然从其氨基酸顺序看它是进化上最为保守的蛋白质中的一种，但其同义替换的速率却非常高。对表 4－1 中的基因来说，非同义替换的平均速率为每年每非同义位点 0.85×10^{-9} 替换。同义替换的平均速率为每年每同义位点 4.6×10^{-9} 替换，即为非同义替换平均速率的 5 倍。

非编码区

来自非编码区的数据远不如来自编码区的数据丰富，所以目前只做过有限的比较分析工作。（注意，要估出某一序列中的替换速率，我们必须至少有来自两个物种的数据。）因为大多数已发表的序列为 mRNA，它们不含内含子和侧区域，所以，其 5′和 3′不翻译区是唯一能进行仔细研究的非编码区. 表 4－2 列出了根据人与啮齿类比较得到的 16 种基因中这两个区域的替换速率。在这两个区域中不同基因间的速率变化都非常大，但这种变异可能很大程度上代表了抽样所造成的影响，因为这两区域通常都非常短。在几乎所有基因中，5′和 3′不翻译区中的速率都低于四重简并位点上的替换速率（即，其上所有可能的核苷酸替换都是同义替换的位点）。5′和 3′不翻译区的平均速率分别为每年 1.96×10^{-9} 和 2.10×10^{-9} 替换，它们都约为四重简并位点上的平均速率，每年 3.55×10^{-9} 替换的 60%。

表 4－2　根据人与小鼠或大鼠的基因间比较，得到的为蛋白质编码的基因的 5′及 3′不翻译区和四重简并位点上的核苷酸替换速率[a]

基　因	5′不翻译区		3′不翻译区		四重简并位点	
	L[b]	速率	L	速率	L	速率
ACTH	99	1.87±0.41	97	2.32±0.49	275	2.78±0.34
醛缩酶 A	124	1.08±0.26	154	1.73±0.32	195	3.16±0.48
载脂蛋白 A－Ⅳ	83	3.06±0.68	134	1.73±0.33	160	3.38±0.50
载脂蛋白 E	23	1.27±0.69	84	1.70±0.42	153	4.00±0.60
Na,K-ATP 酶 β	118	2.45±0.45	1117	0.57±0.06	118	2.87±0.54
肌酸激酶 M	70	1.71±0.46	168	1.79±0.30	178	2.81±0.41
α-胎蛋白	47	3.64±1.13	144	2.79±0.49	225	4.14±0.54
α-珠蛋白	34	1.56±0.65	90	2.21±0.50	81	4.47±0.98
β-珠蛋白	50	1.30±0.46	126	2.85±0.49	78	2.42±0.56
甘油醛-3-磷酸脱氢酶	70	1.34±0.38	121	1.74±0.36	170	2.43±0.39
生长激素	21	1.79±0.85	91	1.83±0.41	83	3.82±0.78
胰岛素	56	2.92±0.80	53	3.09±0.81	62	4.19±1.00
白细胞中介素 I	59	1.09±0.38	1046	2.02±0.14	105	2.97±0.60
乳酸脱氢酶 A	95	2.79±0.55	470	2.48±0.23	152	3.64±0.60
金属巯基组氨酸三甲基内盐 Ⅱ	61	1.88±0.52	111	2.57±0.48	23	2.37±1.00
甲状旁腺素	84	1.79±0.43	228	2.21±0.30	38	3.85±1.21
平均[c]		1.96(0.78)		2.10(0.61)		3.33(0.69)

a. 速率以每 10^9 年每位点替换数为单位。　b. L＝位点数。

c. 平均指算术平均，括号内的值为标准偏差，都是用所有基因的值算出的。

假基因（pseudogenes）是一些由功能基因派生，但由于发生了阻止其正常表达的突变而退化成无功能的 DNA 序列（第六章和第七章）。由于它们不受功能限制，所以，它们可以期望以较高的速率进化。表 4－3 列出了乳牛和山羊的 $\psi\beta^{X}$ 和 $\psi\beta^{Z}$ 假基因中替换速率间的比较，以及 β－和 γ－珠蛋白基因中非编码区和四重简并位点上的速率间的比较。这些假基因中的速率事实上略高于其他区域中的速

率。看来这一点对假基因来说是普遍成立的，尽管目前有关假基因的资料尚属有限。

表 4－3　乳牛和山羊的β－和γ－珠蛋白基因间的分岐，以及β－珠蛋白假基因间的分岐

统计量	β-和γ-珠蛋白基因[a]						假基因
	5′FL	5′UT	四重简并	内含子	3′UT	3′FL	
百分比分岐	5.3	4.0	8.6	8.1	8.8	8.0	9.1
标准误差	1.2	2.0	2.5	0.7	2.2	1.5	0.9

a. FL＝侧区域；UT＝不翻译区域；四重简并＝四重简并位点。

图 4－2 中，我们对基因的不同区域中，以及假基因中的替换速率进行了比较。关于 5′和 3′不翻译

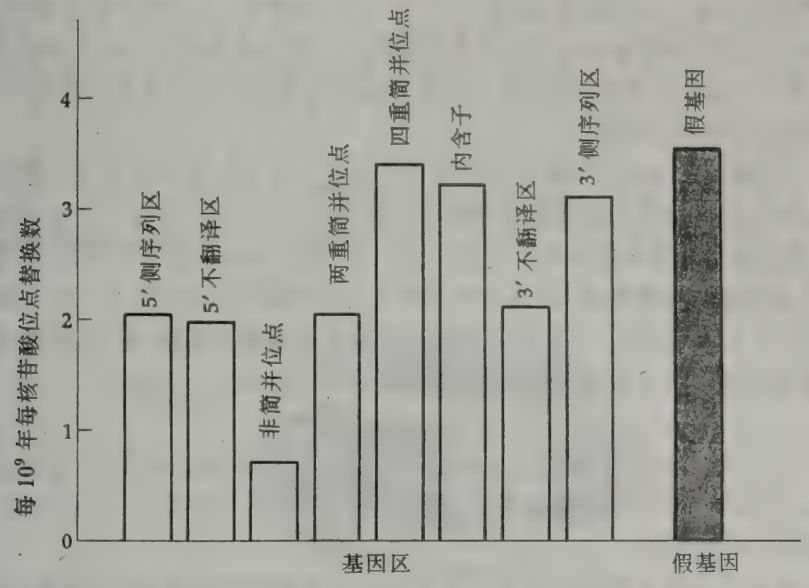


图 4－2　基因的不同部位中以及假基因中的替换平均速率。

区域，非简并位点，两重简并位点和四重简并位点的速率，都是将表 4－2 中所列基因加以平均后的平均速率。5′侧区域的速率，通过假定该速率与四重简并位点上的速率之比为 5.3/8.6（即由表 4－3 得出的值），和四重简并位点上的平均速率为每年 3.33×10^{-9} 替换（表 4－2）而算出。内含子的速率，3′侧区域的速率以及假基因的速率也按同样方式算出。由于以有限的资料为基础而作出的估计已经够多了，又由于一个区域中的速率会因基因的不同而有差异，所以，图 4－2 中展示的速率可能对任何一个具体的基因都是不适用的，但它却提供了不同 DNA 区域中的替换速率间一个大致的、一般性的比较。有了这种思想准备，我们将看到，一个基因中的替换速率以四重简并位点上的为最高，内含子中和 3′侧区域中要略低一些，3′不翻译区域，5′侧区域，5′不翻译区域和两重简并位点有中等大小，而非简并位点上的为最低。平均下来假基因有最高的替换速率，虽然它只比一个功能基因的四重简并位点上的速率稍高一些。

4.2　替换速率变异的原因

为了推理出 DNA 区域间替换速率出现变异的原因，我们应注意到，替换速率是由两个因子所决定的：(1)突变率和(2)一个突变的固定概率（第二章）。后者又与该突变是有利的、中性的还是有害的有关。由于突变率看来在一个基因内变化不大而在不同基因间则可能变化较大，所以，我们将对一个基因的不同区域间的速率变异和不同基因间的速率变异分别讨论。

不同基因区域间的变异

我们首先考虑一个基因中同义位点和非同义位点间的大差异。由于一个基因内同义位点与非同义位点上的突变率应该相同，或者至少是非常相似，所以，替换速率上的差异就可归因于两种不同类

型位点间纯洁化选择的强度上的差异。这可用分子进化的中性学说来理解(第二章)。结果会导入氨基酸替换的突变比同义的改变对该蛋白质的功能造成有害影响的机会要高。所以,绝大多数非同义突变都将受纯洁化选择而从群体中清除。其结果将使非同义位点上的替换速率降低。相比之下,同义的改变有较高的机会是中性的,而它们中在群体中固定的也要多些。

当然,非同义替换可能有幸使蛋白质的功能得到改善。然而,如果有利选择在该蛋白质的进化中起主要作用的话,则非同义替换的速率应该超过同义替换的速率。事实上,在某些免疫球蛋白基因里,决定互补性的区域(CDRs,又以高可变区著称)中非同义的速率高于同义的速率。这种较高的速率已经归因于对抗体多样性的超显性选择(Tanaka 和 Nei,1989)。不过,当考虑的是整个免疫球蛋白基因时,非同义的速率仍然大大低于同义的速率(表 4－1)。这个结果指出,即使在免疫球蛋白中,大多数非同义突变也是不利的,并且将从群体中清除。休斯和根井(Hughes 和 Nei,1989)曾报导在主组织相容性复合体基因的某些区域中有类似情形,即非同义替换的速率超过同义替换的速率。他们把非同义替换有更高的速率归因于超显性选择。

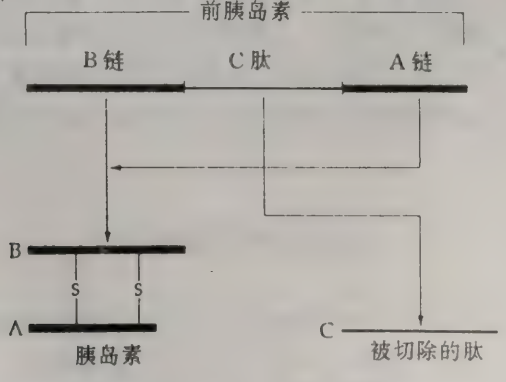
一个基因中同义的和非同义的速率间的对比证明了分子进化中一个众所周知的原则,即,对某一大分子的功能限制越强,则其进化的速率就越缓慢。木村(Kimura,1983)曾用一个简单模型将此原则表达成一个公式。假定某一分子中所有突变的某一部分 f_0,是选择中性或近中性的,而其余的则是有害的(有利突变被假定仅极偶然地出现,以至其相对频率的有效数字为零,而且它们对整个分子进化的速率没有多大贡献)。如果我们用 v_T 表示每单位时间每位点的总突变率,那么,中性突变的突变率即为 $v_0=v_Tf_0$。根据分子进化的中性学说,替换速率为 $K=v_0$(第二章)。因此,

$$K = v_T f_0 \tag{4.2}$$

在任一给定基因内,该 v_T 值可假定对同义位点和非同义位点都是相同的。然而,f_0 值则是同义位点的比非同义位点的高,所以前者要比后者进化得快。虽然该模型是过于简单了,但它对解释不同 DNA 区域间速率上的差异却很有帮助。

依上述模型看,最高的速率预期应出现在一个没有任何功能的序列中,由于没有功能,所以它里面的所有突变都是中性的(即 $f_0=1$)。事实上,假基因看起来的确有最高的核苷酸替换速率(表 4－3 和图 4－2)。5′和 3′不翻译区有比编码区中的同义替换更低的替换速率,这一观察事实进一步支持解释问题的中性路线,因为这些区域含有关于转录起始和终止的信号。

在一个蛋白质内,有不同结构和功能的区域看来受着有差别的功能限制,并以不同速率进化着。胰岛素原为此提供了一个极好的例子。它由 A,B 和 C 三个片段组成(图 4－3),片段 C 位于分子的中间,并在活性激素(胰岛素)形成期间被除去,即胰岛素是由余下的 A 和 B 两个片段所构成的。片段 C



进化速率 0.13×10^{-9}/位点/年　　进化速率 0.97×10^{-9}/位点/年

图 4－3　为有功能的胰岛素(A 和 B 链)和 C 肽编码的 DNA 区域中核苷酸替换速率间的比较。成熟的胰岛素分子由一条 A 链和一条 B 链,通过二硫键(s)联结而成。自 Kimura(1983)修改而成。

对胰岛素的激素活性不起任何作用,而被认为只对产生该激素的正确三级结构有促进效果。结果,为 C 片段编码的区域的非同义替换速率,为 A 链和 B 链编码的区域的平均非同义替换速率的 7 倍多(图 4－3)。然而,C 片段上一定仍受着相当程度的限制,因为该区域中的非同义替换速率还是比较低的,它与 β－珠蛋白中的相应速率大致相当(表 4－1)。

基因间的变异

为了对基因间非同义替换速率方面的大变异作出解释，我们必须再次考虑这样两个可能的肇事者：突变率和选择强度。不同的基因有相同的突变率，这样的假定在这种情况下可能不能成立，因为基因组的不同区域可能有着不同的突变倾向。沃尔夫等(Wolfe 等，1989a)曾提出，哺乳动物细胞核基因组的不同区域，在突变率方面的差异相互间以一个数值为 2 的因子来区别。然而，不同基因组区域间突变率上出现 2 倍的差异，甚至不能算作造成非同义替换速率方面将近 1000 倍的出入的部分原因。所以，决定非同义替换速率的最重要因素看来还是选择强度，它又转而由功能限制所决定。

为了说明功能限制的效应，让我们考虑一下载脂蛋白和组蛋白 3，它们有着差异显著的非同义替换速率。载脂蛋白是脊椎动物血液中各种脂类的主要载体，而它们的脂类结合区大部分由疏水性残基所组成。对来自哺乳纲各目的载脂蛋白的顺序比较分析表明，该区域内由一个疏水性氨基酸(比如缬氨酸、亮氨酸)去替换另一个疏水性氨基酸，这在许多位点上都是可以接受的(Luo 等，1989)。这种不太严格的结构要求可用来解释为什么这些基因中的非同义的速率会相当高(表 4－1)。

处于另一个极端的是组蛋白 3。因为组蛋白 3 中的大多数氨基酸在核小体(图 4－4)形成时，将直接与 DNA 或其他核心组蛋白相互作用，所以，可以合理地假定，只有很少几种可能的替换能在不妨

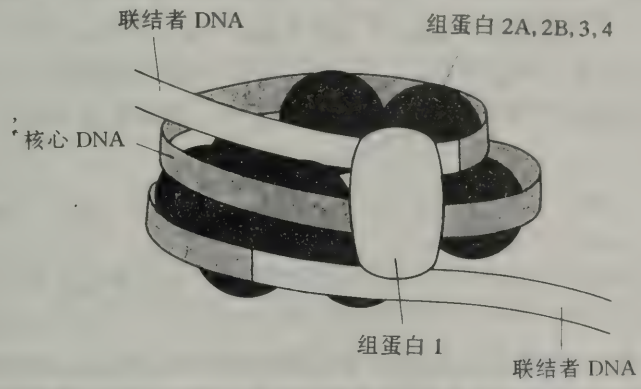


图 4－4　一个核小体的示意图。DNA 双螺旋(黑色带)围着核心组蛋白(组蛋白 2A、2B、3 和 4 各二个)缠绕。组蛋白 1(淡灰色)与该核心粒子的外部和联结者 DNA(白色带)相结合。自 Stryer (1988)修改而成

碍该蛋白质功能的条件下发生。此外，组蛋白 3 还必须维持其严格的致密性和高度碱性，这对与酸性的 DNA 分子相互作用是必要的。结果，组蛋白 3 对大多数分子变化都绝不容忍。事实上，这种蛋白质是已知的进化最为缓慢的蛋白质之一，比载脂蛋白要慢 1000 多倍。

同义替换的速率为什么也会因基因而异还不太清楚。出现这种变异可能有两个原因。首先，基因组的不同区域间突变率可能是不同的，因而同义替换速率上的变异可能简单地反映出基因所处的染色体位置(Wolfe 等，1989a)。这种可能性因这样的事实而得到进一步的支持，即，真核生物的基因组是由被称为同质段的明显地以 GC 为内容的片段所构成的，这些片段可能是独立复制的因而可能表现出不同的突变率(第八章)。第二个原因可能是，在某些基因中，并不是所有密码子都是在适合度上等价的。结果，有些同义替换可能会受到选择的排斥。这种纯洁化选择就会在基因间产生同义替换速率方面的变异。然而，虽然纯洁化选择已被证明能影响同义替换的速率，能影响细菌、酵母和果蝇的基因组中同义密码子的使用模式，但现在还不清楚这类选择是否在哺乳类中发生作用(见第 55 页)。

还有一种现象也曾受到注意，即一个基因中的同义替换速率和非同义替换速率间存在一种正相关(Graur，1985；Li 等，1985b)。若假定突变率随基因而变(因此有些基因的同义和非同义的替换速率将都很高)，或者假定同义位置上的选择大小受邻近的非同义位置上核苷酸组成的影响，则该现象即可得到解释(Ticher 和 Graur，1989)。

4.3　一个正选择例子：乳牛和叶猴的溶菌酶

如前面的章节所讨论的那样，基因组的绝大多数基因和非基因区域中核苷酸替换的速率和模式，

都可以通过①突变输入，②中性或近中性等位基因的随机遗传漂变，和③排斥有害等位基因的纯洁化选择，这三方面的结合来加以解释。然而，在溶菌酶的例子中，对有利突变的正选择曾被证明在某些哺乳类谱系中起作用。

前肠发酵消化曾在有胎盘哺乳类的进化中独立地两次出现，一次在反刍动物（例如乳牛）中，另一次在疣猴类（例如叶猴）中。在这两种情况中，对别的哺乳动物来说通常不在胃中分泌产生的溶菌酶，它能补充进来，以消化在前肠执行发酵任务的细菌的细胞壁。斯图尔特和威尔逊（Stewart 和 Wilson，1987）曾对来自乳牛、叶猴、狒狒、人、大鼠、马和鸡的溶菌酶进行过氨基酸顺序比较（表 4—4）。他们注意到，乳牛和叶猴间有 4 个独特地共有的氨基酸。对这一观察结果有两种可能的解释。第一种：有可能乳牛在进化上与叶猴的亲缘关系比与马的更近，于是这些独特地共有的氨基酸，只代表出现在它们的共同祖先中未发生改变的氨基酸顺序。乳牛和叶猴间系统发生关系更近的假定已知是错误的。另一种，这些在该两物种中独特地共有的氨基酸，可能是独立地发生在两个谱系中的一系列平行替换的结果。事实上，将氨基酸替换的顺序重建后，斯图尔特和威尔逊（Stewart and Wilson，1987）发现在乳牛和叶猴谱系中有 7 个平行或趋同替换（图 4—5）。

表 4—4　不同物种的溶菌酶间顺序的成对比较[a]

物种	物种					
	叶猴	狒狒	人	大鼠	乳牛	马
叶猴		14	18	38	32	65
狒狒	0		14	33	39	65
人	0	1		37	41	64
大鼠	0	1	0		55	64
乳牛	4	0	0	0		71
马	0	0	0	0	1	

自 Stewart 和 Wilson(1987)。

a. 对角线上的数为物种间氨基酸的差异数，而对角线下的数为物种间独特地共有的残基数。

而且，已经确定，这些替换中有些对溶菌酶在低 pH 值下更好地行使功能有贡献，象在反刍动物

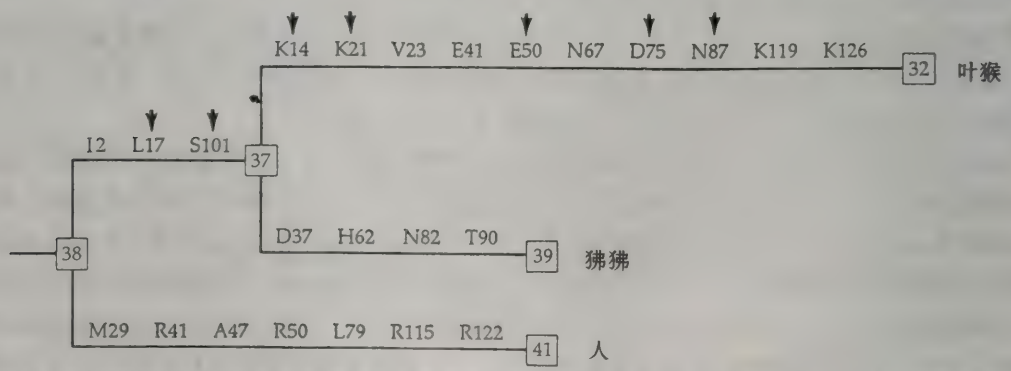


图 4—5　乳牛和叶猴溶菌酶中的平行或趋同氨基酸替代。谱系的长度与沿该谱系发生的氨基酸替代数成正比。每次替代用替代后氨基酸的一字母缩写（表 1—1）表示，其后续数字表示替代发生的位置。箭头指出了叶猴中发生的 7 次替代，它们以与乳牛谱系平行或趋同的形式发生。与乳牛胃溶菌酶（没有画出）的氨基酸差异数写在方块之中。自 Stewart 和 Wilson，(1987)。

消化系统中发 现的那些酶。相反，叶猴和乳牛溶菌酶在高 pH 值下都不如人溶菌酶有效。最后，看来可以不出什么差错地推论，我们这里处理的是一个有利替换在不同进化路线中平行地发生，表现出对类似的选择因子平行地适应的例子。

4.4　分子钟

在对来自不同物种的血红蛋白和细胞色素 c 的蛋白质顺序比较研究中，朱克坎德尔和波林

(Zuckerkandl 和 Pauling,1962,1965)以及马戈利阿什(Margoliash,1963)首次注意到,这些蛋白质中的氨基酸替换速率在不同的哺乳类谱系中近似相同。因此朱克坎德尔和波林(Zuckerkandl 和 Pauling,1965)提出,对任何给定的蛋白质而言,分子进化的速率在所有谱系中都对时间近似地恒定,换言之,就是存在着一种分子钟(molecular clock)。这一提议马上激起了人们对将大分子用于进化研究的极大兴趣。事实上,如果蛋白质是以恒定的速率进化着的,那么,它们将可用于决定物种分岐的年代,并用来重建生物间的系统发育关系。这将与通过测定放射性元素的衰变来决定地质年代类似。

分子钟假说也激起了大量的争论。例如,经典进化论学者们就反对这种说法,因为进化速率恒定的说法与在表型和生理学水平上进化速率的变化无常对不上号。当速率恒定假说用于估计人与非洲猿间的分岐时间,得到一个 500 万年的估值(Sarich 和 Wilson,1967)时,该假说受到了特别强烈的反对。因为,在古生物学家中占统治地位的观点是,人和猿的分岐至少应在 1500 万年前,两者差得太远。许多分子进化科学家们也对分子钟假说的正确性提出了异议。特别是古德曼(Goodman,1981) 以及他的同事们(Czelusniak 等,1982)。他们认为,进化速率常常在基因重复之后出现加速,而蛋白质顺序在适应性辐射的年代里进化要快得多。例如,他们主张,在基因重复使 α 和 β 血红蛋白分开之后出现了极高的氨基酸替换速率,而这种高替换速率则是由于改善血红蛋白功能的有利突变所造成的。

虽然速率恒定假说一直是有争议的,但它已广泛用于分岐时间的估计和系统发育树的重建中(Nei,1975; Wilson 等,1977)。所以,分子钟假说的正确性在分子进化中是一个生死攸关的问题。近几年中 DNA 顺序数据的迅速积累为检验该假说提供了一个全新的机会。用这类数据与用蛋白质顺序数据相比可使我们更近地检验该假说,而与 DNA—DNA 杂交数据和免疫距离数据相比,则可得到更直接的解释。

相对速率测验

关于分子钟假说的争论常常引起有关物种分岐年代的异议。为了避免这一问题,萨里奇和威尔逊(Sarich 和 Wilson,1973)提出了一种不需要知道分岐年代的检验法。该检验法称相对速率测验(relative-rate test),如图 4—6 所示。假定我们要比较谱系 A 和 B 中的速率。于是,我们用第 3 个物种 C 作为参照物。我们应该确定,该参照物种的分岐过程发生得比物种 A 和 B 间的分岐更早。例如,为了比较人和马来猩猩谱系中的速率,我们用一种猴作为参照物。

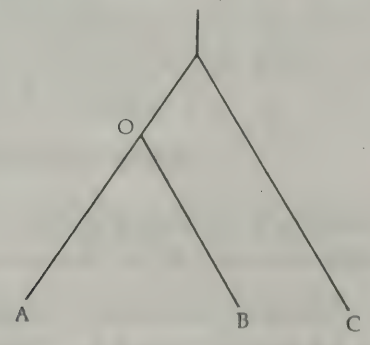


图 4—6 用于相对速率测验的系统树。O 表示物种 A 和 B 的共同祖先。

从图 4—6 很容易看出,物种 A 和 C 间的替换数 K_{AC} 等于从点 O 到点 A 发生的替换数(K_{OA})和从点 O 到点 C 发生的替换数(K_{OC})之和。即,

$$K_{AC} = K_{OA} + K_{OC} \tag{4.3a}$$

类似地,有:

$$K_{BC} = K_{OB} + K_{OC} \tag{4.3b}$$

和

$$K_{AB} = K_{OA} + K_{OB} \tag{4.3c}$$

因为 K_{AC},K_{BC} 和 K_{AB} 能从核苷酸顺序直接估出(第三章),所以,我们可以很容易地解出这 3 个方程,找到 K_{OA},K_{OB} 和 K_{OC} 的值:

$$K_{OA} = \frac{K_{AC} + K_{AB} - K_{BC}}{2} \tag{4.4a}$$

$$K_{OB} = \frac{K_{AB} + K_{BC} - K_{AC}}{2} \tag{4.4b}$$

$$K_{OC} = \frac{K_{AC} + K_{BC} - K_{AB}}{2} \tag{4.4c}$$

现在我们可以通过K_{OA}和K_{OB}的值，来决定替换速率在谱系A和谱系B中是否相等。自物种A和B最后地共有一个共同祖先以来所经过的时间，定义为对两个谱系是相等的。所以，按分子钟假说，K_{OA}和K_{OB}应该相等，即，$K_{OA}-K_{OB}=0$。从等式4.3a和4.3b可以看出，$K_{OA}-K_{OB}=K_{AC}-K_{CB}$。故而，我们可以从$K_{AC}$和$K_{BC}$来直接比较A和B中的替换速率。

小鼠和大鼠中接近相等的速率

表4—5展示了用相对速率测验法进行的小鼠和大鼠中同义替换速率的比较。表中的物种A皆指小鼠，而物种B则全表示大鼠。因此，若$K_{AC}-K_{BC}$的值为一正号，则表示小鼠中的速率高于大鼠中的，而若为负号则指示实际情况正好相反。

表4—5　小鼠(物种A)和大鼠(物种B)间每100位点的同义替换数差异(K_{AC}-K_{BC})[a]

基因	L	K_{AB}	K_{AC}	K_{BC}	$K_{AC}-K_{BC}$
载脂蛋白E	201	7.4	61.3	59.5	1.8±5.3
肌动蛋白α	249	17.9	58.2	59.1	−0.9±4.8
肌动蛋白β	233	19.7	50.1	45.1	5.0±4.6
Thy-1抗原	116	19.3	51.8	57.3	−5.5±6.9
乳酸脱氢酶A	219	30.9	80.4	80.3	0.1±8.2
糖蛋白激素，α亚基	58	30.8	97.7	84.3	13.4±18.5
胰岛素样生长因子Ⅱ	130	4.8	37.0	40.9	−3.9±2.8
心房钠尿因子	107	20.4	69.7	57.4	12.3±8.3
生长激素	124	14.1	80.9	79.2	1.7±7.7
甲状腺球蛋白β	90	25.7	77.4	92.7	−15.3±12.9
鸦片黑素皮质激素原	154	21.4	61.5	52.7	8.8±6.5
醛缩酶A	184	15.4	57.5	63.3	−5.8±5.3
肌酸激酶M	251	17.2	48.6	52.2	−3.6±4.3
金属巯基组氨酸三甲基内盐Ⅱ	35	19.0	45.5	36.7	8.8±10.2
总计	2187	19.0	59.8	59.4	0.4±1.5

自Li等(1987a)

a. L—受比较位点数。K_{ij}—物种i和j间每100位点的替换数。人为参照物种(c)，但肌酸激酶M为一例外，该行数据是用兔的顺序作为参照物得出的。

由于受资料限制，我们用作参照物的是人或兔的顺序，而不是从与小鼠和大鼠的亲缘关系更近的物种，象仓鼠或豚鼠中得到的顺序。结果，$K_{CA}-K_{BC}$的估值表现出较大的统计误差(表4—5)。不过，小鼠和大鼠中的替换速率接近相等，这一点是相当明显的。换句话说，当将所有顺序一起考虑时，该速率差接近于0。关于这两个物种中非同义替换的速率，可得出同样的结论(Li等，1987a)。

人中的速率低于猴的速率

根据免疫距离和蛋白质顺序数据，古德曼(Goodman，1961)及其同事们(Goodman等，1971)提出，人科动物(人和猿)自它们从远古时代与猴分离后，出现了速率减缓。然而，威尔逊等(Wilson等，1977)反驳说，减缓是一种人为现象，是用了对人—猿分歧时间的错误估值的结果。他们用免疫距离数

据和蛋白质顺序数据做了两次相对速率测验，结论是，没有任何表明人科动物速率减缓的证据。

DNA 顺序数据可对上述争论给出一个更好的解决。在表 4－6 中，相对速率测验法被用于比较人谱系和古世界猴谱系的核苷酸替换速率。在所有检验中，谱系 B 是人的谱系，而谱系 A 为猴的谱系。谱系 C 是参照物种（见表下的注释）。因此，速率差（$K_{AC}-K_{BC}$）为正号意味着人谱系曾较慢地进化着，而为负号则意思相反。

表 4－6　古世界猴谱系（A）和人谱系（B）间每 100 位点的核苷酸差异数（$K_{AC}-K_{BC}$）[a]

顺序	位点数	K_{AB}	$K_{AC}-K_{BC}$
η-珠蛋白假基因	2000	7.4	2.1±0.7**
同义位点			
β-珠蛋白	71	8.9	2.8±5.6
载脂蛋白 A-I	158	7.9	−5.3±4.8
促红细胞生成素	145	11.2	5.1±5.9
α_1-抗胰蛋白酶	140	10.9	6.7±6.8
胰岛素	84	18.6	−7.5±7.2
内含子			
δ-珠蛋白	601	4.7	3.4±1.4
不翻译区和侧区域			
β-珠蛋白	179	4.6	1.2±1.7
δ-珠蛋白	172	8.8	6.1±3.2
总计	3550	6.7	2.3±0.6**

a. 所用参照物种为枭猴（η－珠蛋白假基因），狐猴（β－和 δ 珠蛋白），小鼠或大鼠（促红细胞生成素，载脂蛋白 A－Ⅰ，和 α1－抗胰蛋白酶）和狗（胰岛素）。

** 在 1% 水平与 0 差异显著。

我们注意到，9 个所用的顺序中只有 2 个为负号。速率上的差异即使在只用 η 假基因时也是很显著的。若将所有顺序一起考虑，则 $K_{AC}-K_{BC}=2.3\%$，而 $K_{AB}=6.7\%$。因此，古世界猴谱系中的 K 值（K_{OA}）为 $(6.7\%+2.3\%)/2=4.5\%$，人谱系中的 K 值（K_{OB}）只有 $6.7\%-4.5\%=2.2\%$，表明猴谱系以 $4.5/2.2\approx2$ 倍于人谱系的速率更快地进化着。

啮齿类中的速率高于灵长类中的速率

吴和李（Wu 和 Li，1985）曾用相对速率测验法来比较啮齿类谱系和人类谱系中的替换速率，参照物则用偶蹄类或食肉类谱系。他们的结论是，啮齿类谱系同义替换的速率约为人类谱系的 2 倍。不过要注意，速率上的估计差异指长期内（即从啮齿类—灵长类分歧到现在的时期）的平均值。由于在发生分歧的那一时期和其后的短时期里两谱系中的替换速率应是相似的，所以，速率差异一定是随时间而增加的。因此，以上估计的速率差异值可能是一个低估值。要想知道更近时期的速率差异，我们需要分别估出啮齿动物内和灵长动物内的替换速率。

表 4－7 展示了灵长类和啮齿类中的替换速率的比较。如果我们假定人—黑猩猩的分歧发生在七百万年前，而人—古世界猴的分歧发生在二千五百万年前，则人和黑猩猩的顺序间平均速率为每年每位点 1.3×10^{-9} 替换，而人与古世界猴的顺序间则为每年每位点 2.2×10^{-9} 替换。这一结果与古世界猴谱系以 2 倍于人谱系的速率进化的结论是一致的。如果我们假定小鼠—大鼠的分歧发生在一千五百万年前，则小鼠和大鼠的顺序间平均速率为每年每位点 7.9×10^{-9} 替换。因此，啮齿类中的速率可能是较高等的灵长类中的速率的 4 到 6 倍。虽然关于所用的分歧时间还不是很可靠，但啮齿类顺序的进化比灵长类顺序要快得多，这一点是很明显的。

表 4—7　灵长类和啮齿类中每年每位点同义替换的平均速率[a]

物种对	位点数	百分比分歧	替换速率($\times10^{-9}$)
人对黑猩猩	921	1.9	1.3(0.9—1.9)[b]
人对古世界猴	998	11.0	2.2(1.8—2.8)
小鼠对大鼠	3886	23.7	7.9(3.9—11.8)

自 Li 等(1987a)

a. 所用分岐时间，人—黑猩猩为 7(5—10)百万年前，人—古世界猴为 25(20—30)百万年前，小鼠—大鼠为 15(10—30)百万年前。

b. 括号内的值为从分岐时间的上限估值和下限估值得到的速率估值构成的范围。

不同进化谱系间替换速率上出现变异的原因

猴的替换速率高于人的以及啮齿类的替换速率高于灵长类的，这也许能用所谓世代时间效应(generation-time effect)来加以解释(Kohne，1970)。啮齿类的世代时间比人的要短得多，所以，如果在这些生物间每世代的种系复制没有很大差别的话，则每年的种系 DNA 复制的次数在啮齿类中就可能比在人类中要高许多倍。因为突变大多在 DNA 复制的过程期间积累，所以，复制的周期数越多，则突变错误也将发生得越多。这一因素也许能在很大程度上解释啮齿类的替换速率高于人类的替换速率的现象。类似地，猴有比人短的世代时间，所以，应该预期它将有较高的替换速率。

替换速率上的差异也可部分地归因于 DNA 修复系统的效率方面的差异(Britten，1986)。已有的有限资料表明，啮齿类有比人类效率低的 DNA 修复系统，因而，在每一复制周期中将积累更多突变。

以上结果不应拿来作为不存在分子钟的证据。我们注意到，替换速率上的差异是在具有很不相同的世代时间的生物间被观察到的。当具有相近的世代时间的生物，象小鼠和大鼠进行比较时，速率恒定规律表现得相当明显。所以，虽说没有一个关于所有哺乳类的全球性时钟，但关于许多亲缘关系较近的物种类群的地方性时钟也许是存在的。

4.5　细胞器 DNA 中的替换速率

与细胞核基因组相比，细胞器基因组要小得多，也更容易进行实验研究。而且，在哺乳动物线粒体基因组中替换速率特别高(Brown 等，1979)。这一发现激起了人们对细胞器 DNA 的进化问题的更大兴趣。

哺乳动物线粒体基因组由一个环状、双链 DNA 组成。长为 15,000—17,000 碱基对(bp)，近似地相当于最小的动物细胞核基因组的 1/10,000。它只含有单一的(即非重复的)序列：13 个为蛋白质编码的基因，2 个 rRNA 基因，22 个 tRNA 基因和一个调控区，后者含有复制和转录的起始位点。该基因组在结构上是非常稳定的，这一点，从不同种的哺乳动物间其基因组大小变异不大即可看出。

与其成鲜明对照的是，植物的线粒体基因组却展现出较大的结构变异性。它们经历了频繁的重排、重复和缺失(Palmer，1985)。为此，基因组大小在 40,000bp 到 2,500,000bp 的范围内变化。植物中的线粒体基因组可能是线状的，也可能是环状的，而在许多情况中，遗传信息被分割成相互独立的 DNA 分子，后者被称为亚基因组环。植物线粒体的编码内容还没有全部确定；不过，我们确已知道，有 3 个确定 rRNA 的基因、数目还不清楚的 tRNA 基因和大约 15 个—30 个为蛋白质编码的基因，其中有些已被鉴别出来了。(在植物线粒体基因组中结构基因可能以多重拷贝出现。)目前，尽管植物线粒体基因组在大小上有很大变异性，但在编码内容上却没有表现出性质变异的迹象。

维管植物的叶绿体基因组是环状的，大小在 120,000 到 220,000bp 的范围内变化，平均大小为 150,000bp(Palmer，1985)。尽管在大小上有如此大的变异，但该基因组已知在结构上是稳定的。烟草(*Nicotiana tabacum*)的叶绿体基因组已经被完全定序了(Shinozaki 等，1986)。它是一个环状分子，长 155,844bp 含有 37 个 tRNA 基因(其中 8 个含有单内含子)，8 个 rRNA 基因，和 45 个为蛋白质编码的基因(其中 5 个含有单内含子，其中 2 个含有两内含子)。两条链都用于编码。*Nicotiana tabacum* 的

叶绿体基因组还含有 59 个功能不知的外加开读框架，其中 2 个则插进了内含子。

哺乳动物线粒体基因中同义替换的速率已估出，为每年每同义位点 5.7×10^{-8} 替换(Brown 等，1982)。这大约是细胞核中为蛋白质编码的基因的同义替换值的 10 倍。非同义替换的速率在 13 个为蛋白质编码的基因中变化很大，但通常都比细胞核基因的平均非同义替换速率大得多。哺乳动物线粒体中这些高替换速率的原因，看来是相对于细胞核而言它有较高的突变率。高突变率则是由于(a)线粒体中的 DNA 复制过程保真度低，(b)缺乏修复机制或修复机制效率极差，和(c)诱变剂浓度高(例如超氧化物基团 O_2^-)，后者是线粒体执行代谢功能的结果。另一方面，作用在非同义突变上的纯洁化选择的强度，看来与作用在细胞核基因上的属同一数量级。

根据几个基因顺序或限制酶图谱资料进行的早期研究指出，叶绿体基因有比哺乳动物细胞核基因低的核苷酸替换速率(Curtis 和 Clegg，1984；Palmer，1985)，用核苷酸替换表示则植物线粒体 DNA 进化缓慢，虽然它频繁地经历着顺序重排(Palmer 和 Hebron，1987)。这些结果近来被更广泛的 DNA 顺序分析所证实(Wolfe 等，1987，1989)。

表 4－8 展示了高等植物的这 3 种基因组中替换速率的比较。每非同义位点的平均替换数(K_A)在叶绿体和线粒体基因组中是相似的，但每同义位点的平均替换数(K_S)却很不相同，在单子叶植物与双子叶植物间比较，叶绿体基因组中的 K_S 几乎是线粒体基因组中的 3 倍；而在玉米与小麦或大麦间比较则前者是后者的 6 倍。以下，我们将采用前一个比值，因为它是根据更大的数据组而得到的。植物细胞核基因中的平均同义替换速率约为叶绿体基因的 4 倍。于是植物线粒体、叶绿体和细胞核基因中的同义替换速率，近似地呈 1∶3∶12 这样的比例。

表 4－8 植物叶绿体、线粒体和细胞核基因中核苷酸替换速率的比较[a]

基因组	K_S	L_S	K_A	L_A
单子叶与双子叶植物间的比较				
叶绿体基因	0.58±0.02	4177	0.05±0.00	14421
线粒体基因	0.21±0.01	1219	0.04±0.00	4380
玉米与小麦或大麦间的比较				
细胞核基因	0.71±0.04	1475	0.06±0.00	5098
叶绿体基因	0.17±0.01	2068	0.01±0.00	7001
线粒体基因	0.03±0.01	413	0.01±0.00	1526

自 Wolfe 等，(1987，1989b)

a. K_S：每同义位点的替换数；K_A：每非同义位点的替换数；L_S：同义位点数；L_A：非同义位点数。

如果我们把玉米与小麦间的分歧时间取为 50－70 百万年(Stebbins，1981；Chao 等，1984)，那么，表 4－8 中关于细胞核的数据则表现出一个每年每位点 $5.1-7.1\times10^{-9}$ 替换的平均同义速率。这与在哺乳类细胞核基因中看到的同义替换速率(表 4－1)相似。

有趣的是，细胞器的基因组中核苷酸替换的速率与结构变化的速率无关。在哺乳类中，以核苷酸替换表示的线粒体 DNA 的进化非常迅速，但其基因的空间排列和基因组的大小却在各物种间保持稳定。相反，植物的线粒体基因组经历了频繁的结构变化，但其核苷酸替换速率却极低。在叶绿体 DNA 中，核苷酸替换速率和结构进化都很缓慢。替换速率和结构进化速率间无相关性，这表示两个过程是独立地进行的。

4.6 假基因中的核苷酸替换模式

因为点突变是 DNA 序列进化中最重要的因素之一，所以分子进化学家们长期以来一直对决定自发突变的模式(pattern of spontaneous mutation)怀有兴趣(例如，Beale 和 Lehmann，1965；Zuckerkandl 等，1971)。该模式可以被当作一个标准，用于推论：任一给定 DNA 序列中核苷酸间相互变换的观察频率与在无选择下，即在选择中性(selective neutrality)下，预期的值究竟偏离多远。

研究点突变模式的途径之一，是检验不受选择限制的DNA区域中的替换模式。假基因在这里特别有用。由于它们无功能，所以，所有发生在假基因中的突变都是选择中性的，且以相同的概率在群体中固定。于是，假基因中的核苷酸替换模式预期将反映出自发点突变的模式。

图4－7表示推论假基因序列中的核苷酸替换的一种简单方法(Gojobori等，1982；Li等，1984)。

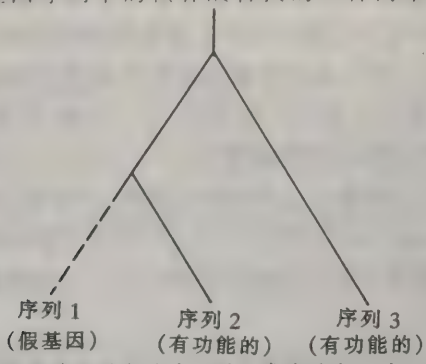


图4－7 推论假基因序列中核苷酸替换模式的系统树。虚线的含义是无功能。

序列1是一个假基因，序列2是其有功能的对应物，来自同一物种，而序列3则是假基因出现前即发生分岐的功能序列。假定在某一核苷酸位点上，序列1和2分别各为A和G。那么，我们可以假定，若序列3中该位点上为G则假基因序列中该核苷酸从G变为A，但若序列3中该位点上为A则序列2中该核苷酸从A变成G。不过，如果序列3中的是T或C，则我们不能决定变化的方向，而若出现这种情况则该位点将从比较中排除。由于假基因中的替换速率通常大大高于其同源的功能基因中的速率，所以，一个基因与一个假基因间核苷酸顺序上的差异，在绝大多数情况下都被认为是假基因中的改变所造成的。

表4－9中的矩阵代表从13种哺乳类假基因序列推论出的替换的联合模式。矩阵中的每一个项f_{ij}，都代表一个随机序列(即四种碱基以相同的频率出现其中的序列)中每100次替换里碱基从i变成j的期望次数。例如，$f_{AT}=4.7$，表示所有替换的4.7%为从A变成T。

表4－9 假基因中的替换模式[a]

从	到 A	到 T	到 C	到 G	行总计
A	—	4.7±1.3 (5.3±1.4)	5.0±0.7 (5.6±0.8)	9.4±1.3 (10.3±1.4)	19.1 (21.2)
T	4.4±1.1 (4.8±1.1)	—	8.2±1.3 (9.2±1.3)	3.3±1.2 (3.6±1.3)	15.9 (17.6)
C	6.5±1.1 (7.1±1.3)	21.0±2.1 (18.2±2.3)	—	4.2±0.5 (4.2±0.6)	31.7 (29.5)
G	20.7±2.2 (18.6±1.9)	7.2±1.1 (7.7±1.3)	5.3±1.0 (5.5±1.3)	—	33.2 (31.8)
列总计	31.6 (30.5)	32.9 (31.2)	18.5 (20.3)	16.9 (18.1)	

自Gojobori等，(1982)和Li等，(1984)

a. 表中项为以13种哺乳类假基因序列为根据推论出的，碱基从i变为j的百分数(f_{ij})。括号中的值是把所有CG二核苷酸从比较中排除后得到的。

我们注意到，突变的方向是非随机的。例如，A变成G比变成T或C更常发生。从右上角到左下角的对角线上的4个元是转换的f_{ij}值，其余的8个元代表颠换。所有转换，特别是C→T和G→A，都比颠换更常发生。转换的相对频率之和为59.2%(若CG二核苷酸被排除则为54.4%，见下)。我们注意到在随机突变下转换的期望比例仅33%，因为只有4种转换却有8种颠换。该观察比例几乎是在随机突变下预期值的两倍。

我们也注意到，有些核苷酸比另一些要更容易突变。在表4－9的最后一列，我们列出了从A、T、

C 和 G 突变成别的核苷酸的相对频率。如果 4 种核苷酸都有相同的突变性，则我们应期望该列中的每一个元都有一个 25% 的值。实际上，我们看到，G 以 33.2% 的相对频率突变（即，G 是一种高可突变的核苷酸），而 T 则以 15.9% 的相对频率突变（即它达不到那种可变程度）。在表 4－9 的最后一行，我们列出了所有经突变而变成 A、T、C 和 G 的相对频率。我们注意到，所有突变的 64.5% 是变成 A 或 T 的，而随机过程期望的值应为 50%。由于 C 和 G 有一种频繁地变成 A 或 T 的倾向，又由于 A 和 T 不如 C 和 G 那样可突变，所以，假基因预期应变得富含 A 和 T。这对其他一些不受功能限制的非编码区应该也是成立的。事实上，非编码区普遍发现是富 AT 的。

表 4－9 中的结果是根据有意义的链，即未被转录的链而得出的。所以，从 G 到 A 的变化实际上意味着一个 G:C 对被一个 A:T 对所取代。这种情况的出现，可以是有意义的链中 G 突变成 A 的结果，也可能是与前者互补的链中 C 突变成 T 的结果。类似地，从 C 到 T 的变化，也可能是在一条链中 C 突变成 T 或在另一条链中 G 突变成 A 的结果。如果两条链间突变的模式没有差别，那么我们应有 $f_{GA}=f_{CT}$。类似地，我们应能得到 $f_{AG}=f_{TC}$，$f_{AT}=f_{TA}$，$f_{AC}=f_{TG}$，$f_{CA}=f_{GT}$ 和 $f_{CG}=f_{GC}$。这些等式仅近似地成立，且事实上这两条链间的突变模式可能存在较小的不对称性。这种不对称性可能是由于 DNA 复制期间，先导链和滞后链间在复制机制方面有差异所造成（Wu 和 Maeda，1987）。

已知从 C 到 T 的转换，除了碱基误配以外，还可能是从甲基化了的 C 残基经脱氨而变成 T 残基，这样一种转变过程而实现的（Coulonder 等，1978；Razin 和 Riggs，1980）。该作用将提高 C:G→T:A 和 G:C→A:T，即 f_{CT} 和 f_{GA} 的频率。由于脊椎动物 DNA 中约 90% 的甲基化了的 C 残基发生在 5′－CG－3′二核苷酸中（Razin 和 Riggs，1980），所以，该效应将主要以 CG 二核苷酸变成 TG 或 CA 的形式表现出来。一个基因变成假基因后，这类变化将不再受任何功能限制，因而，如果在基因的沉默化（silencing）（即失去功能）发生前 CG 的频率相对而言较高，则它能为 C→T 和 G→A 转换作出显著贡献。对看来曾在这假基因的祖先序列中出现过 CG 二核苷酸的那些位点予以排除，由此而得到的替换模式在表 4－9 的括号中给出。此模式也许更适合于：预测一个长期不受功能限制的序列（例如一个内含子的某些部分）中的突变模式，因为在这样的序列中将只有少量 CG 二核苷酸存在。排除 CG 二核苷酸后得到的模式有点不同于不经排除而得到的模式。特别地，4 种转换间的相对频率差异显著性略有降低，而各颠换的相对频率，除 G→C 和 C→G 外，则略有升高。

4.7 同义密码子的非随机应用

由于遗传密码的简并，20 种氨基酸中大多数都是由一个以上的密码子编码的（第一章）。因为同义突变不造成氨基酸顺序中的任何变化，且因为自然选择被认为主要在蛋白质水平上起作用，所以，同义突变曾被当作选择上呈中性的突变的候选者（Kimura，1968；King 和 Jukes，1969）。然而，若所有同义突变事实上都是选择中性的，那么，为同一个氨基酸编码的同义密码子就应该以多少有点相同的频率应用。不幸的是，随着 DNA 顺序资料的积累，逐渐表明同义密码子的应用，在原核生物和真核生物的基因中都显然是非随机的（Grantham 等，1980）。事实上，在许多酵母的基因和大肠杆菌的基因中，应用上的偏斜是极显著的。例如，在大肠杆菌（*Escherichia coli*）外膜蛋白Ⅱ（*omp* A）中的 23 个亮氨酸残基里，有 21 个由密码子 CUG 编码，尽管为亮氨酸编码的还有 5 种密码子。这种偏斜不能用非随机突变来解释。如何解释这种广泛存在的密码子非随机应用现象，成了一个有争议的问题，好在对此问题看来已出现了一些明确的答案。

有助于理解非随机应用现象的一个观察事实是，一个生物中或有亲缘关系的物种中的基因，一般表现出对同义密码子的选取有同样的模式（Grantham 等，1980）。于是，哺乳动物、大肠杆菌和酵母的基因被归为不同的密码子应用类型。格兰瑟姆等（Grantham 等，1980）因此而提出了基因组假说（genome hypothesis）。按其假说，任何给定基因组中的基因在同义密码子的选取方面都采用同样的编码策略，即在密码子应用上的偏斜是物种特异的。基因组假说被证明一般说来是正确的，虽然在一个基因组的不同基因间密码子应用有着相当大的异质性（见下面）。

大肠杆菌和酵母中密码子应用的研究，极大地增加了我们对影响同义密码子选取的因素的认识。

波斯特等(Post 等,1979)发现,大肠杆菌核糖体蛋白基因,优先应用于被含量最多的 tRNA 种类识别的同义密码子。他们认为,这种偏向是自然选择的结果,因为应用由含量最多的 tRNA 种类翻译的密码子,将会增加翻译的效率和精确性。他们的发现曾激励池村(Ikemura,1981,1982)去收集有关大肠杆菌和酵母 *Saccharomyces cerevisiae*(酿酒酵母)中各 tRNA 种类的相对丰度的资料。他证明,在这两个物种中,一个基因中同义密码子的相对频率与识别它们的 tRNA 种类的相对丰度间存在着正相关。对于高度地表达的基因而言,这种相关非常强。例如,在大肠杆菌中 4 种亮氨酸 tRNA 里含量最多的是 $tRNA_1^{Leu}$,它识别 CUG 密码子,而 *ompA* 基因也主要用这种密码子为亮氨酸编码(见上面)。

图 4—8 表示 6 个亮氨酸密码子的频率和识别它们的 tRNA 的相对丰度间的对应关系。在大肠杆

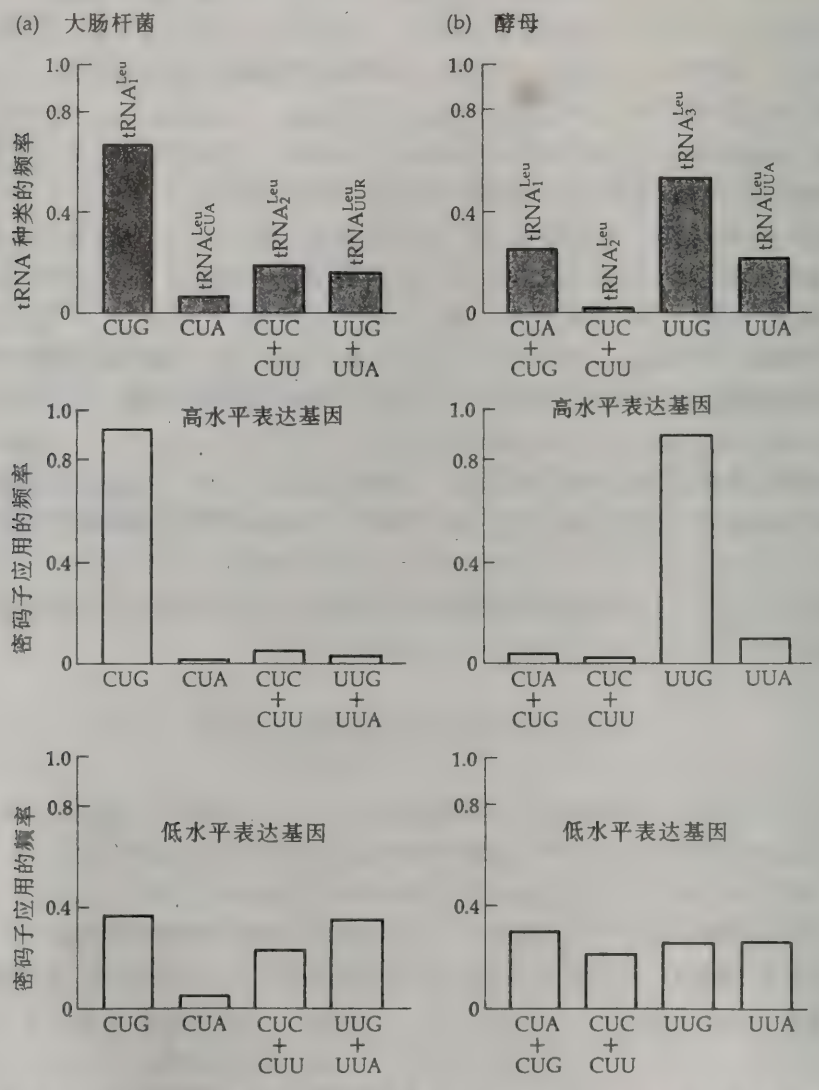


图 4—8 关于亮氨酸的密码子应用的相对频率(白柱体)和相应的识别 tRNA 种类的相对丰度(黑柱体)间关系的图解说明。(a)在大肠杆菌中和(b)在 *Sacharomyces cerevisiae* 中加号(例如 *E. coli* 中密码 CUC 和 CUU 间的符号)表示这类密码子对中的密码子都是由同一种 tRNA 来识别的(例如,*E. coli* 中 CUC 和 CUU 都由 $tRNA_2^{Leu}$ 来识别)。

菌中,$tRNA_1^{Leu}$ 是含量最多的 tRNA 种类,而事实上,在高水平表达的基因中,CUG(由这种 tRNA 识别的密码子)的应用比另外 5 种密码子要频繁得多。另一方面,在酵母中,含量最丰富的亮氨酸 tRNA 种类是 $tRNA_3^{Leu}$,而被这种 tRNA 识别的密码子(UUG)也是数量上占优势的密码子。对比之下,在以较低水平表达的基因中,tRNA 丰度和各密码子间的对应在这两个物种里都要弱得多(图 4—8)。

在决定高水平表达的基因中密码子的应用模式方面,翻译效率的重要性已得到以下观察的进一步支持(Ikemura,1981)。已知密码子-反密码子配对在第 3 位上出现摇摆(wobbling)。例如,反密码子的第 1 位上的 U 既可与 A 也可与 G 配对。类似地,G 既可与 C 也可与 U 配对。但是,反密码子第 1 位

上的C则只能与密码子第3位上的G配对，以及A只能与U配对。摇摆还可能通过这样的事件来实现：有些tRNA在第1反密码位置上含有经修饰过的碱基，而这类tRNA能识别一种以上的密码子。例如，次黄嘌呤（一种经过修饰的腺嘌呤）可与U、C、A三种碱基中的任何一种配对。有趣的是，大多数能识别一种以上密码子的tRNA，都表现出对其中的某一种有不同的偏爱。例如，反密码子的摇摆位置上的4－硫尿嘧啶（S^4U），可以识别密码子摇摆位置上的A和G；然而，与以G结尾的密码子相比，它对以A结尾的密码子表现出明显的偏爱。这种偏爱在高度地表达的基因中应会反映出来。大肠杆菌中两个为赖氨酸编码的密码子是由一种tRNA识别的，该tRNA分子在反密码子的摇摆位置上有S^4U，而事实上，在大肠杆菌的*omp* A基因中，19个赖氨酸密码子里15个是AAA，只有4个是AAG。

表4－10列出了由夏普等（Sharp等，1988）广泛收集的密码子应用资料中的一部分。对每一组同义密码子来说，如果应用机会均等，则每种密码子的相对频率应该是1。然而，大多数情况下显然并非如此。而且，在大肠杆菌和酵母这两个物种中，密码子应用偏斜都是在高水平表达的基因中比在低水平表达的基因中更严重。对此差异的一个简单解释是，在高水平表达的基因中对翻译效率和精度的选择要强一些，所以密码子应用偏斜也就显著一些。另一方面，在低水平表达的基因中选择相对而言较弱，所以，该应用模式主要受选择压力和随机遗传漂变的影响，因而偏斜程度也低一些（Sharp和Li，1986）。

表4－10　4个物种中的密码子应用[a]

氨基酸	密码子	*Escherichia Coli*		*Saccharomyces Cerevisiae*		*Drosophila Melanogaster*		人	
		高	低	高	低	高	低	G+C	A+T
Leu	UUA	0.06	1.24	0.49	1.49	0.03	0.62	0.05	0.99
	UUG	0.07	0.87	5.34	1.48	0.69	1.05	0.31	1.01
	CUU	0.13	0.72	0.02	073	0.25	0.80	0.20	1.26
	CUC	0.17	0.65	0.00	0.51	0.72	0.90	1.42	0.80
	CUA	0.04	0.31	0.15	0.95	0.06	0.60	0.15	0.57
	CUG	5.54	2.20	0.02	0.84	4.25	2.04	3.88	1.38
Val	GUU	2.41	1.09	2.07	1.13	0.56	0.74	0.09	1.32
	GUC	0.08	0.99	1.91	0.76	1.59	0.93	1.03	0.69
	GUA	1.12	0.63	0.00	1.18	0.06	0.53	0.11	0.80
	GUG	0.40	1.29	0.02	0.93	1.79	1.80	2.78	1.19
Ile	AUU	0.48	1.38	1.26	1.29	0.74	1.27	0.45	1.60
	AUC	2.51	1.12	1.74	0.66	2.26	0.95	2.43	0.76
	AUA	0.01	0.50	.00	1.05	0.00	0.78	0.12	0.64
Phe	UUU	0.34	1.33	0.19	1.38	0.12	0.86	0.27	1.20
	UUC	1.66	0.67	1.81	0.62	1.88	1.14	1.73	0.80
Met	AUG	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00

自Sharp等，(1988)

a. 对每一同义密码子组来说，相对频率之和等于该组中的密码子数。例如，亮氨酸有6个密码子，所以关于这6个密码子的相对频率之和即应为6。在均等地应用下，一组中每一密码子的相对频率应为1，所以各值与1的偏差指示应用上偏斜的程度。"高"和"低"表示以高水平表达和以低水平表达的基因。对人来说，"G+C"意味着高GC区，而"A+T"意味着高AT区。

总之，在大肠杆菌和酵母中，同义密码子的选取受着RNA可用性和其他与翻译效率有关的因素的限制。这些限制结果将以纯洁化选择表现出来，从而减缓了同义替换的速率（Ikemura，1981；Kimura，1983）。事实上，已经得到证明的是，肠道细菌的基因中同义替换的速率与密码子应用的偏斜程度呈负相关（Sharp和Li，1986）。因此，同义密码子的非随机应用现象不能被当作反对分子进化的中性学说的证据，因为它可用选择限制越强结果进化速率就越低这一原理来加以解释（见第45页和Kimura，1983）。

表4－10还展示出，果蝇中的密码子偏斜在高水平表达的基因中比在低水平表达的基因中要严重得多，这指出对翻译效率的选择在决定这种生物里同义密码子的选取方面也起重要作用。

在许多人的基因中，密码子倾向于以G或C结尾(即在第3位置上有较高的GC含量)，而在另一些基因中却有较低的第3位置GC含量。不过，有几种原因可用来说明为什么该偏斜可能与基因表达的水平无关。首先，α－和β－珠蛋白基因在密码子的第3位上有不同的GC含量(分别为高含量和低含量)，但它们在同样的组织(红细胞)中以近似相等的量表达，因而它们应该有相同的表达水平。其次，在鸡的基因中，密码子应用的频率与tRNA的可用性无关(Ouenzar等，1988)，虽然该观察也许不能直接地用于人的基因。最后，第3密码子位置上的GC含量与侧区域中和内含子中的GC水平有很强的相关性(第八章，Bernardi和Bernardi，1985；Aota和Ikemura，1986)。例如，α－珠蛋白基因GC含量高且它位于高GC区域，β－珠蛋白基因GC含量低而它位于低GC区域(第八章，Bernardi等，1985)。于是，看来人基因中的密码子应用偏斜，极大地因含有该基因的区域中GC的含量而决定。正如将要在第八章讨论的那样，一个区域中的GC含量是由自然选择还是由突变偏斜所决定，这仍是一个有争议的问题。不过，由于一个基因中第3密码子位置上的GC含量倾向于高于其周围区域中的含量(第八章；Aota和Ikemura，1986)，所以，有可能人的基因中密码子应用模式受到某种程度的自然选择的影响。要得到对影响人中密码子应用的因素的更多知识，还需要做进一步的研究。

习题

1. 等式4.1的分母为什么是2T不是T？

2. 图4－9表示来自橄榄狒狒(*Papio anubis*)和马来猩猩(*Pongo pygmaeus*)的Θ1－珠蛋白基因的第1和第2外显子的DNA顺序。用一参数模型分别算出3个密码子位置中每一个的每位点替换数。哪一个位置进化得最快？为什么？

b: ATG GCG CTG TCC GCG GAG GAC CGG GCGGCT GTG CGC GCC CTG
o: ATG CGC CTG TCC GCG GAG GAC CGG GCGCTG GTG CGT GCC CTG

b: TGG AAG AAA CTG GGA AGC AAT GTT GGCGTC TAT GCT ACT GAG
o: TGG AAG AAG CTG GGC AGC AAC GTC GGCGTC TAC ACG ACA GAG

b: GCC CTG GAG AGG ACC TTC CTG GCT TTCCCC GCC ACG AAG ACC
o: GCC CTG GAG AGG ACC TTC CTG GCC TTCCCC GCA ACG AAG ACC

b: TAC TTC TCC CAC CTA GAC CTG AGC CCCGGC TCC GCC CAG GTT
o: TAC TTC TCC CAC CTG GAC CTG AGC CCCGGC TCC TCA CAG GTC

b: AGA GCA CAC GGC CAG AAG GTG GCG GACGCG CTG AGC CTC GCC
o: AGA GCC CAC GGC CAG AAG GTG GCG GACGCG CTG AGC CTC GCC

b: GTG GAG CGC CTA GAC GAC CTA CCC CGCGCG CTG TCC GCT CTG
o: GTG GAG CGC CTG GAC GAC CTA CCC CACGCG CTG TCC GCG CTG

b: AGC CAT CTG CAC GCT TGC CAG CTG CGAGTG GAC CCA GCT AAC
o: AGC CAC CTG CAC GCG TGC CAG CTG CGAGTG GAC CCG GCC AGC

b: TTC CCG
o: TTC CAG

图4－9 来自橄榄狒狒(b)和马来猩猩(o)的Θ1－珠蛋白基因中，外显子1和2的DNA顺序。资料取自Shaw等，(1987)和Marks等，(1986)。

3. 图4－10表示来自橄榄狒狒和马来猩猩的Θ1－珠蛋白基因中第1个内含子的DNA顺序。用(a)一参数模型，和(b)两参数模型，算出替换数。这两个估计有差别吗？将此结果与你在习题2中得到的结果相比较，那么，该内含子的进化比外显子中3个密码子位置上的进化是快还是慢？

b: TGCGGCGAGGCTGGGCGCCCCCGCCCTCCGGGGCCCTGCCTCCCCAAGCC
o: TGCGGCGAGGCTGGGCGCCCCCGCCCCC-AGGGCCCTCCCTCCCCAAGCC

b: CCCCGGACGCGCCTCACCGCCGTTCCTCTCGCAG
o: CCCCGGACTCGCCTCACCCACGTTCCTCTCGCAG

图 4—10 来自橄榄狒狒(b)和马来猩猩(o)的 θ1—珠蛋白基因中第 1 个内含子的 DNA 顺序。一个裂缝用-标出。资料取自 Shaw 等,(1987)和 Marks 等,(1986)

4. 图 4—11 表示来自仓鼠、大鼠和小鼠的核仁素基因中的第 1 个内含子的部分 DNA 顺序。以仓鼠的顺序作为参照物,用相对速率检验法来决定,大鼠谱系和小鼠谱系间替换速率上是否存在差异。

m: GTAAGAGGCCTGGCGCGCCGACGCGGACGACTAGGCCTGCTTTCGGAGGG
r: GTAATAGGCCTGACGCGCGAACACGGACGACTAGGCCTGCTTTCTGAGAG
h: GTGAGAGGCCTCGCGCGCGCCGACGGACGGACGGGCCTGCTTTCTGAGGG

m: GCGCGCGCGCCGTCGCGGAGGGGAGGAGGGCTTGCGCGCAATCCCGGGCG
r: GCGCGCGCGCCGTCGCGGAGGGGAGGAGGGCCTGCGCACAGTCCCGGGCG
h: GCGCGCGCGCGGTCGCTCAGGGGAGGAGGGCCTGCGCGCAATCCCGGGCG

m: CGTTCGAGGGCGCCAGCTGGGGAACTCTCGCGCGACTAGCGGGAGGTCTC
r: CGTTCGAGGGCGCATGCTGGGGAAGTCTCGCGCGACTAGCGGAGGGTCTC
h: CGTTCGAGGGCGCATGCTGGGGAAGTCTCGCGCGACTAGCGGAGGGTCTC

图 4—11 来自小鼠(m)、大鼠(r)和仓鼠(h)的核仁素基因中第 1 个内含子的部分 DNA 顺序。裂缝(缺失和插入)已被略去。资料取自 Bourbon 等,(1988)

5. 艾滋(AIDS)病毒的两个品系用 WMJ1 和 WMJ2 表示,在 1984 年 10 月 3 日和 1985 年 1 月 15 日从一个两岁的孩子身上分离出来(Hahn 等,1986)。这个孩子假定只受过一次感染(由她的母亲在围产期传给)。这两个分离物间外壳蛋白(*env*)基因中每同义位点的同义替换数为 0.0164(Li 等,1988)。(a)假定 WMJ2 直接从 WMJ1 进化而来并假定这两个顺序在 1984 年 10 月 3 日分开,求同义替换速率的最大估值。(b)假定这两个品系在被感染时即开始分歧,并假定它们已独立地进化了两年,求该基因同义替换速率的最小估值。这些替换速率比对哺乳动物的基因加以平均的同义替换速率(表 4—1)要快多少?

6. 从大肠杆菌、酵母和人中各找出一个完整的 cDNA 或基因顺序。对每一种基因编一个密码子应用表(即列出每种密码子在基因中被使用的次数)。那么,该密码子应用是否偏斜?在哪方面偏斜?密码子应用模式在这 3 种基因中是否相似?如果不是,则差异如何?用 X^2 检验法去判断,在每一种基因中缬氨酸密码子的应用与该族密码子机会均等地应用的偏差是否具有统计学意义。

后继阅读文献

Ikemura. T. 1985. Codon usage and tRNA content in unicellular and multicellusar organisms. Mol. Biol. Evol. 2: 13—34

Kimura. M. 1983. *The Neutral Theory of Molecular Evolution*. Cambridge University Press. Cambridge.

MacIntyre. R. J. (ed.). 1985. *Molecular Evolutionary Genetics*. Plenum. New York.

Nei. M. 1987. *Molecular Evolutionary Genetics*. Columbia University Press. New York.

Sharp. P. M. and W. -H. Li. 1986 An evolutionary perspective on synonymous codon usage in unicellular organisms. J. Mol. Evol. 24: 28—38

Steinhauer. D. A. and J. J. Holland. 1987. Rapiel evolution of RNA viruses. Annu. Rev. Microbiol. 41: 409—433.

5 分子系统发育

分子系统发育是用分子生物学技术对生物间的进化关系进行的研究。它是分子进化中的一个领域，近十年来已引起了广泛关注，这主要是因为在许多情况下系统发育关系用任何别的方法都难以估价的缘故。本章的目的是，解释如何用分子数据来构建系统树，以及给出一些例子，以说明对一些存在已久的系统发育问题，分子方法能够提供比传统方法要清晰得多的解答。

5.1 分子数据对系统发育研究的影响

分子系统发育的研究发端于本世纪开始前后，甚至早于孟德尔定律被重新发现的1900年。免疫化学研究表明，血清学交叉反应在亲缘关系较近的生物间比在关系较远的生物间强。这些发现的进化含义被纳托尔(Nutall，1904)用来推论各不同动物类群间的系统发育关系。例如，他断定人类最近缘的亲属是猿，其后，亲缘关系由近及远依次为：古世界猴，新世界猴和原猴。

自十九世纪六十年代后期以来，分子生物学中的各种技术都得到发展，从而开始了将分子数据广泛应用于系统发育的研究。特别是在十九世纪七十年代和十九世纪八十年代，用蛋白质顺序资料研究分子系统发育进展极快。花费不大但便利得多的方法，象蛋白质电泳，DNA－DNA杂交和免疫学方法，虽然不如蛋白质顺序测定来得精确，却也被广泛用于群体间或亲缘关系较近的物种间的系统发育关系的研究中。这些方法的应用也刺激了遗传距离的测定和系统树构建法的发展(见Fitch和Margoliash，1967；Nei，1975；Felsenstein，1988)。

自十九世纪八十年代后期以来的DNA顺序资料的积累，已经对分子系统发育产生了巨大的影响。DNA顺序资料不仅更为丰富，而且比蛋白质顺序资料更容易分析。因此，它们一方面已被用于推断象人与猿那样亲缘关系很近的物种间的系统发育关系(见第72页)，另一方面又被用来研究一些非常古老的进化事件，象线粒体和叶绿体的起源(见第76页)，门和界的分化(Woese，1987)等。将来，DNA定序看来将用于解决系统发育研究中许多长期得不到解决的问题，象细菌和单细胞真核生物间的进化关系(Sogin等，1986，1989)，这用任何传统的进化研究方法都不可能得到解决。事实上，分子数据在进化史的研究方面被证明是极为有用的，藉此也许我们最终能完美地构建出生物界主要类群的系统发育树。

当然，我们也不放弃进化研究的传统手段，象形态学、解剖学、生理学和古生物学等。因为，不同方法可提供互相补充的数据。我们注意到，分类学主要以形态学和解剖学资料为依据，而古生物学信息则是能提供用于进化研究的时间框架的唯一资料。

5.2 系统树

地球上的一切生命形式，不管是现存的还是已经灭绝了的，都有一个共同的起源，它们的祖先可以追溯到大约在40亿年以前生存的一种或几种生物。因此，所有动物、植物、细菌通过祖藉而相互关联。亲缘关系近的生物是由一个较近代的共同祖先传下来的，亲缘关系远的生物则由较远古的共同祖先传下来。系统发育研究的目的是：(1)建立各生物间正确的系谱学联系，和(2)估计各生物自它们从最后一个共同祖先那里分岐以来的分岐的时间。

在系统发育研究中，一组生物类群间的进化关系常用系统树(phylogenetic tree)来图示说明。系统树是一种由节点和分枝组成的图象，其中任何两个邻近的节点都只由一个分枝来联结(图5－1)。节

点(nodes)代表分类学单位,而分枝(branches)则用祖藉和祖先来定义这些单位间的关系。一个树的分枝模式被称为拓扑图(topology)。枝长(branch length)通常代表在该分枝中曾发生过的变化数。由节点表示的分类学单位可以是物种、群体、个体或基因。

在处理系统树时,我们要分清外节点(external nodes)和内节点(internal nodes)。例如,图 5－1a 中,节点 A、B、C、D 和 E 是外节点,其余的都是内节点。外节点代表处于比较中的现存分类学单位,并被命名为操作中的分类学单位(operational taxonomic units)即 OTU。内节点则代表祖先单位。

图 5－1 为处理一个系统树的两种普通方法的示意图。在图 5－1a 中,其分枝是无尺度的(un-

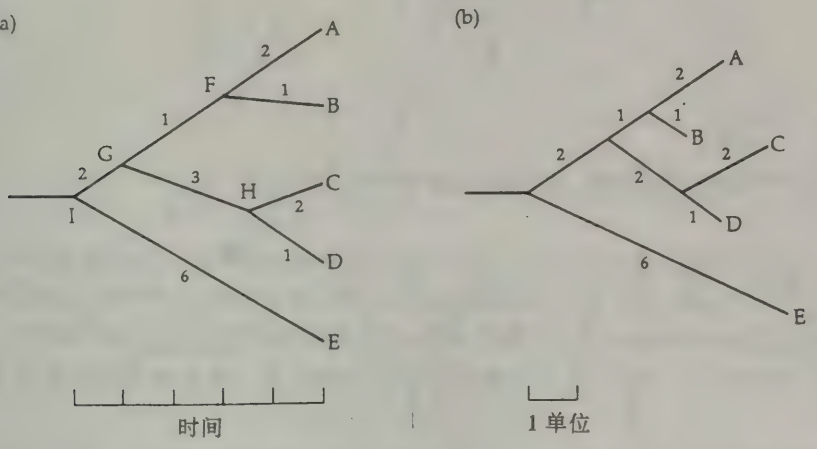


图 5－1 一个关于 5 个 OTU 的系统树的两种可供采用的表示法。(a)无尺度的分枝:现存的 OTU 排成一条线,节点位于与分岐时间成比例的地方。(b)有尺度的分枝:分枝的长度与分子变化数成比例。

scaled),它们的长度与已注明在分枝上的变化数不成比例。这种表示法使我们能将现存的 OTU 排成一条直线,而且在分岐时间已知或已估出时,还可把代表分岐事件的节点按时间尺度来排列。在图 5－1b 中,分岐是有尺度的(scaled),且其长度正比于变化数。

如果任何两个 OTU 间的距离等于将它们联起来的所有分枝的长度之和,则这样的树即被称为是加性的(additive)。例如,若可加性成立,则图 5－1a 中 OTUA 和 C 间的距离应等于 2+1+3+2=8。两个 OTU 间的距离可直接从分子数据(例如 DNA 顺序)算出,而枝长则可按某些规则从 OTU 间距离估出(见第 77 页)。如果在任一核苷酸位点上曾发生过多重替换,则可加性一般不成立(图 3－5)。

若一个节点只有两个直接的后代谱系,则它是两分叉的(bifurcating),若它有两个以上的直接后代谱系则是多分叉的(multifurcating)。为简便起见,我们将只考虑象图 5－1 中那样的两分叉树。

有根树和无根树

系统树可以是有根的(rooted)也可以是无根的(unrooted)(图 5－2)。在有根树中,存在一个被称为根(图 5－2a 中的 R)的特殊节点,由此导向任何别的节点都只有唯一途径。每一途径中的方向与进化时间相对应,而根则是所有正被研究的 OTU 的共同祖先。无根树是一种只将各 OTU 间的关系具体化而未定义进化途径的树(图 5－2b)。

对 3 个物种来说,存在着 3 种可能的有根树,但只有一种无根树(图 5－3)。对 n 个 OTU 而言,两分叉有根树的数目(N_R)由:

$$N_R = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{5.1}$$

给出,其中 n≥2。对 n≥3,两分叉无根树的数目(N_U)为:

$$N_U = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{5.2}$$

注意,n 个 OTU 的可能无根树数等于(n－1)个 OTU 的可能有根树数。OTU 从 2 到 10 的可能有根树

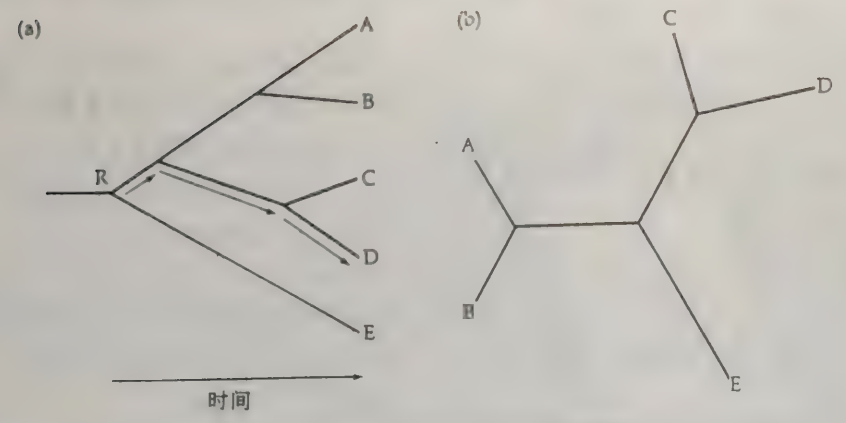


图 5—2 (a)有根系统树和无根系统树 箭头指示从根(R)到 OTUI)的唯一途径

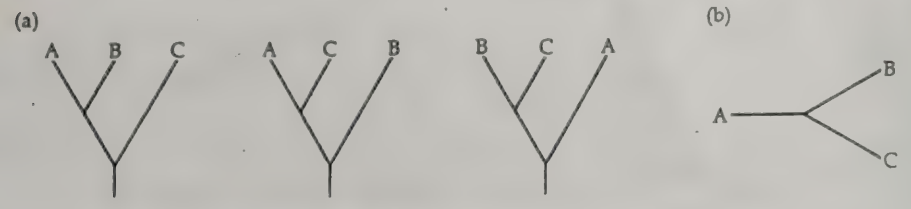


图 5—3 从 3 种 OTU 可能会构建出 3 种不同的有根树(a),但却只有一种无根树(b)。

和无根树的数目列于表 5—1。我们看到,随着 n 的增加 N_U 和 N_R 都极快地增大,而到 10OUT 时已经有 200 多万两分叉无根树和将近 3500 万有根树了。由于这些树中只有一种能正确地表示这些 OTU 间真实的进化关系,所以,当 n 较大时,推论出真实的系统树通常是非常困难的。

表 5—1 对 1—10 OTU 可能的有根树和无根树的数目

OTU 数	有根树数	无根树数
2	1	1
3	3	1
4	15	3
5	105	15
6	954	105
7	10395	954
8	135135	10395
9	2027025	135135
10	34459425	2027025

自 Felsenstein(1987)

真实树和推测树

导致任何类群的 OTU 形成的物种形成事件的秩序,在历史上是唯一的。所以,在用某一给定数的 OTU 建立的所有可能的树中,只有一种能代表真实的进化历史。这样一种系统树称为真实树(true tree)。用某一组数据和某种构树法得到的树称推测树(inferred tree)。推测树可能与真实树等同,也可能与真实树不同。

基因树和物种树

表示一群物种的进化途径的系统树称物种树(species tree)。若系统树是根据来自各物种的一个

基因构成的，则该推测树即为基因树(gene tree)(Nei,1987)。它与物种树有两个方面不同。第一，从两不同物种取样的两基因的分岐可能在时间上早于两物种的分岐(图 5—4)。这将造成高估了枝长的结果，但如果我们处理的是长期进化，在这个过程中由物种内遗传多态性造成的分岐成分可以忽略的话，则还不至于出现严重问题。

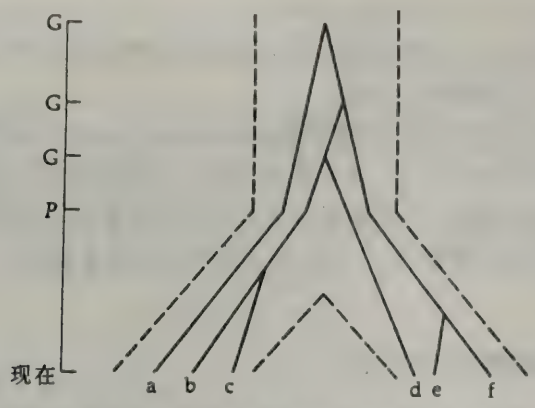


图 5—4　本图表示：若群体在时刻 P 是遗传上多态的，则基因分化(t)通常比群体分化(P)出现得早。结果形成用 a—f 表示的 6 个等位基因的基因进化史用实线表示，群体分岐则以虚线表示。自 Nei(1987)。

基因树的第 2 个问题是，基因树的分枝模式(即它的拓扑图)可能与物种树的不同。图 5—5 展示了这两种树间的 3 种不同的可能关系。基因树(a)与(b)的拓扑图和与其对应的物种树的图等同(例

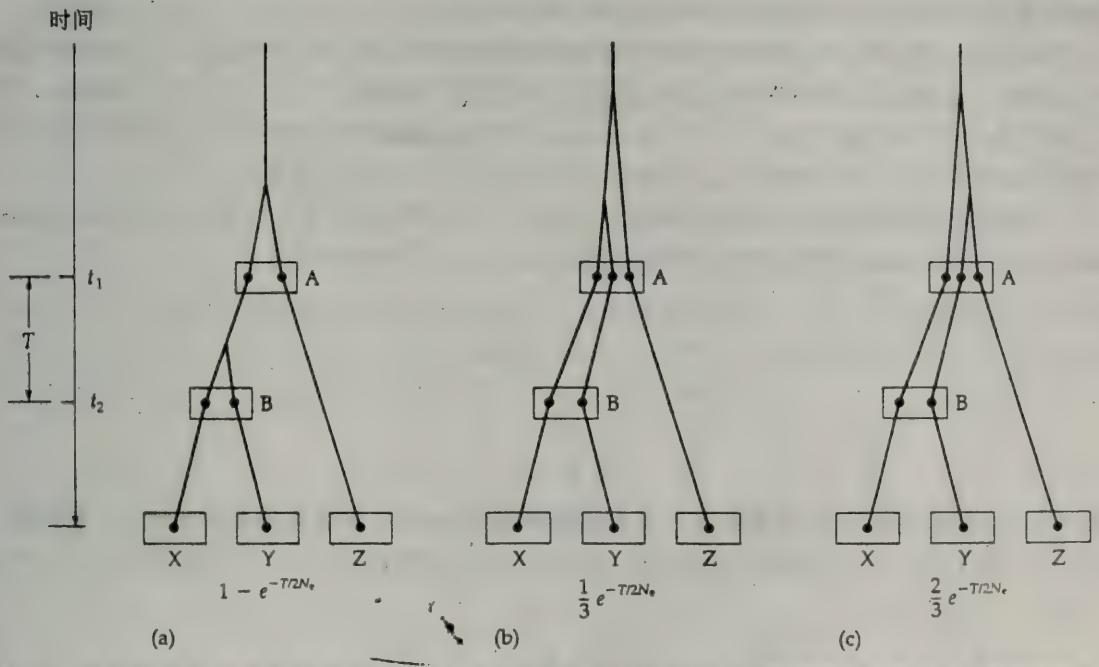


图 5—5　物种树(矩形)和基因树(圆点)间的 3 种可能关系。在(a)和(b)中，物种树的拓扑图与基因树的等同。注意，(a)中基因的分岐时间大致等于群体的分岐时间。而(b)中却不然，基因 X 和 Y 的分岐时间大大早于各自所属的群体的分岐时间。在(c)中，基因树的拓扑图与物种树的不同。对于中性等位基因，每种树出现的概率已在该树的下面列出。t_1 是第一次物种形成事件发生的时间，t_2 是第二次物种形成事件发生的时间。图中 $T=t_1-t_2$，Ne 则是有效群体大小。自 Nei(1987)。

如，X 和 Y 形成一簇)。然而基因树(c)却与真实的物种树不同，因为此时 Y 和 Z 是姐妹类群。当第一次物种分化和第二次物种分化间的时间间隔(t_1-t_2)较短时，得到错误的树(c)的概率将是相当高的，这种情况在确定人、黑猩猩和大猩猩间的系统发育关系时可能确实存在。为了避免得到这种错误类型，在构建系统树时就必须用许多基因。为了避免随机误差也需要大量数据。因为核苷酸替换是随机地发生的，所以有可能会产生随机误差。例如图 5—5a 中，谱系 Z 累积的替换可能会因机遇而比谱系 X 和 Y 的少，尽管它发生分枝的时间要更早一些。

5.3 系统树的构建方法

文献中已有许多构树方法被提出来。有关的详细处理读者可参考史尼斯和索卡尔(Sneath 和 Sokal,1970)、根井(Nei,1987) 以及费尔森施泰因(Felsenstein,1988)等的论述。这里我们讲 4 种在系统发育研究中常被用到的方法。为简单起见,我们考虑核苷酸顺序数据,但所讲到的方法同样也可用于其他类型的分子数据,象氨基酸顺序数据。

以下所讲的方法可分成两种类型:距离矩阵法(distance matrix methods) 和最节省法(maxmum parsimony methods)。在距离矩阵法中,进化距离(通常为分开两分类学单位的核苷酸或氨基酸替换数)按所有可能的分类单位对算出,并用根据距离值间的某些函数关系建立的算法来构建系统树。在最节省法中,应用了特性状态(例如,某一个位点上的核苷酸或氨基酸),而导致这些特性状态的最短途径即被选取为系统树。

不加权算术平均组对法(UPGMA)

不加权算术平均组对法(unweighted pair group method with arithmetic mean)(UPGMA)是一种简单的构树法。它最初本用来构建分类学表型关系图,即能反映出各 OTU 间的表型相似性的树(Sokal 和 Michener,1958;并见第 68 页),但如果进化速率在不同谱系间近似地恒定,以至进化距离与分岐时间之间存在近似的线性关系,则它也可用来构建系统树(Nei,1975)。如果这种关系成立,则象核苷酸(或氨基酸)替换数这样的线性距离尺度就应该被用到了。

UPGMA 法采用连续聚类算法,在此法中局部拓扑关系按相似性的级别来鉴别,而系统树的构建是分步实现的。也就是说,我们首先从所有 OTU 中找出两个最相似的 OTU,并把它们看成一个新 OTU。这样的 OTU 我们称之为复合 OTU(composite OTU)。接着,从这种新的 OTU 群中我们再找出有最大相似性的对子,如此继续下去,直到我们最后仅留下 2 个 OTU 时为止。

为了对此法作出具体说明,让我们来考虑一个包含 4 个 OTU 的例子。成对的进化距离,如朱克斯和坎托(Jukes 和 Cantor,1969)所估计的那样(第三章),由以下矩阵所给出:

	OTU		
OTU	A	B	C
B	d_{AB}		
C	d_{AC}	d_{BC}	
D	d_{AD}	d_{BD}	d_{CD}

在此矩阵中,d_{ij}代表 OTUi 和 j 间的距离。最先被聚类的两个 OTU 是距离最短的那两个。我们假定 d_{AB}有最小值,则 OTUA 和 B 是最先被聚类的,而分枝点则放在距离为 $d_{AB}/2$ 替换的地方(图 5—6a)。

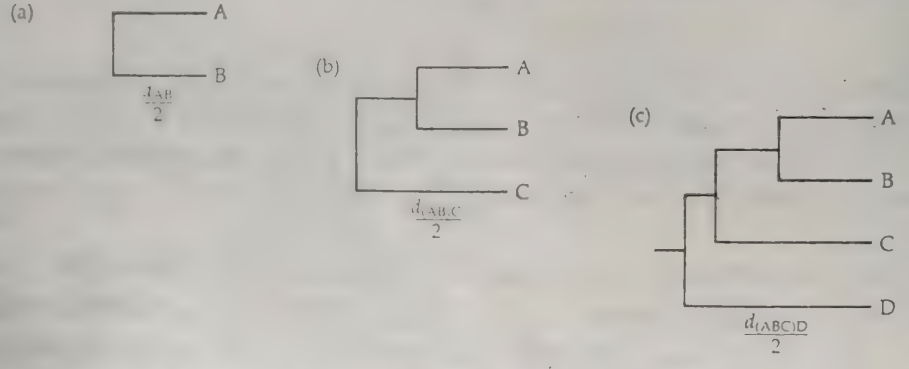


图 5—6 对 4 个 OTU 按 UPGMA 法分步构建系统树的示意图(详见正文)。

第一次聚类后,A 和 B 被看成一个复合 OTU,据此而算出一个新距离矩阵:

	OTU	
OTU	(AB)	C
C	$d_{(AB)C}$	
D	$d_{(AB)D}$	d_{CD}

在此矩阵中，$d_{(AB)C}=(d_{AC}+d_{BC})/2$，$d_{(AB)D}=(d_{AD}+d_{BD})/2$。换言之，一个单 OTU 和一个复合 OTU 间的距离是，该单 OTU 与该复合 OTU 的各组成单 OTU 间距离的平均值。如果 $d_{(AB)C}$ 是新矩阵中的最小距离，那么，OTU C 将加入到复合 OTU(AB)中，且在 $d_{(AB)C}/2$ 处有一个分枝节点(图 5－6b)。

最后一步是将最后剩下的 OTUD 与复合 OTU(ABC)聚类。整个树的根位于 $d_{(ABC)D}/2=[(d_{AD}+d_{BD}+d_{CD})/3]/2$ 处。用 UPGMA 法推测出的最后树如图 5－6c 所示。

在 UPGMA 法中，两个复合 OTU 间的距离用每一复合 OTU 中的所有组成 OTU 间距离的算术平均算出。例如，复合 OTU(ij)和复合 OTU(mn)间的距离为：

$$d_{(ij)(mn)}=\frac{(d_{im}+d_{in}+d_{jm}+d_{jn})}{4} \tag{5.3}$$

在复合 OTU(ijk)和(mn)的例子中，距离为：

$$d_{(ijk)(mn)}=\frac{(d_{im}+d_{in}+d_{jm}+d_{jn}+d_{km}+d_{kn})}{6} \tag{5.4}$$

变形距离法

如果假定速率恒定在各谱系间不成立，则 UPGMA 法可能会给出错误的拓扑图。例如，假定图 5－7a 中的系统树是真实树，而各 OTU 对的进化距离则由以下矩阵给出：

	OTU		
OTU	A	B	C
B	8		
C	7	9	
D	12	14	11

用 UPGMA 法，我们得到的推测树在分枝模式上与真实树不同(图 5－7b)。如，推测树中 OTUA 和 C 被组合在一起了，而在真实树中，A 和 B 才是姐妹 OTU。(注意，此例中可加性不存在。如，A 和 B 间的真实距离为 8，而估出的联结 A 和 B 的分枝，其枝长之和为 3.50＋0.75＋4.25＝8.50)。

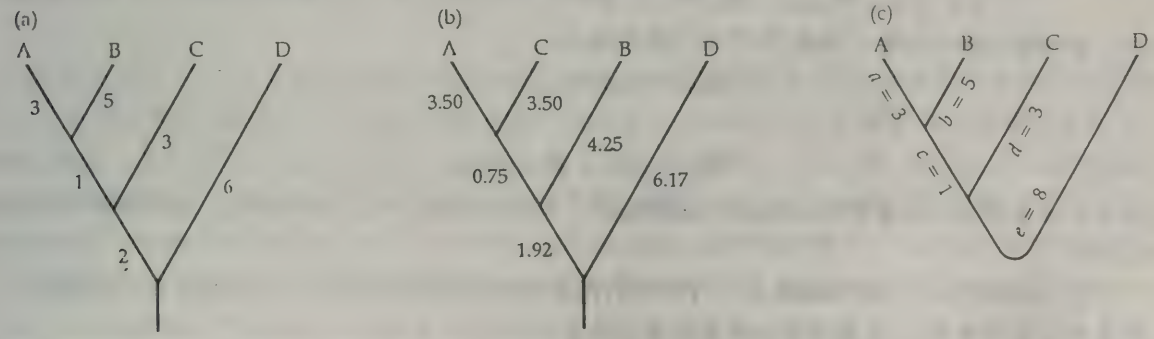


图 5－7 (a)真实的系统树，(b)错误的系统树。此树没有考虑不同分枝的非均一替换速度的概率，运用 UPGMA 方法构造。(c)此树由变形距离法推断得到。根必须在 OTUD 和 OTUA、B 与 C 的共同祖先节点之间，但准确的位点不能确定。

不过，该拓扑图错误也许能用被称为变形距离法(transformed distance method)的校正方法来订正(Farris，1977；Klotz 等，1979)。简单地说，此法是用一个组外单位(outgroup) 作参照物，来对被研究谱系间进化速率不等的情况作一些校正，然后对新得到的距离矩阵应用 UPGMA 法，从而推测出该树的拓扑图。组外单位是一个我们对其有外在认识的 OTU。这些外在认识，如分类学的或古生物学的知识，能清楚地表明，该 OTU 已先于所有其他被研究的 OTU 而从其共同祖先分岐出来。

在现在的例子中，我们假定分类单位 D 对所有别的分类单位来说是一个组外单位。那么 D 即可当作参照物，并用下式来变换距离：

$$d'_{ij} = \frac{d_{ij} - d_{iD} - d_{jD}}{2} + \overline{d_D} \tag{5.5}$$

这里 d'_{ij} 是变形距离，i＝A，B 或 C，$\overline{d_D} = (d_{AD} + d_{BD} + d_{CD})/3$。$\overline{d_D}$ 项的引入是为了保证所有 d'_{ij} 的值都是正的。做此手续是因为，实际上距离不可能是负的。对于有 n 个 OTU（不包括组外单位）的一般情形，$\overline{d_D} = \sum d_{iD}/n$。

在我们的例子里，$\overline{d_D} = \frac{37}{3}$ 而新距离矩阵中关于分类单位 A、B 和 C 的值为

	OTU	
OTU	A	B
B	10/3	
C	13/3	13/3

由于 d'_{AB} 有最小值，所以 A 和 B 最先被聚类在一起，然后，C 再加入到该树中。根据定义，组外单位 OTUD 决定该树的根，因而最后加入该树。这样就给出了正确的拓扑图（图 5－7c）。在上例中，我们仅考虑了 3 个分类单位和 1 个组外单位，但此法可很容易地推广到更多分类单位和/或更多组外单位的场合。

在许多具体情况下，不可能预先决定被研究的分类单位中哪一个是组外单位。为了克服这一困难，两步进行法已被提了出来（Li，1981）。第一步，用 UPGMA 法先推出该树的根。然后，处于根一侧的分类单位被当作参照物（组外单位），对处于根的另一侧的谱系间不相等的进化速率做校正，以后再做相反方向的校正。在我们的例子中，此法也能找出正确的树。

近邻关系法

在一个无根两分叉树中，如果两个 OTU 通过一个内部节点联结，则它们就被说成是近邻（neighbors）。例如，图 5－8a 中，A 和 B 是近邻，C 和 D 也是如此。与之相比，图 5－8b 中，A 和 C 不是近邻，B 和 C 也不是。然而，如果我们把 OTUA 和 B 结合成一个复合 OTU，则复合 OTU（AB）和单 OTUC 就变成了一对新近邻。

现在我们假定，图 5－8a 中展示的树是真实树。那么，若可加性存在则我们应有：

$$d_{AC} + d_{BD} = d_{AD} + d_{BC} = a + b + c + d + 2x = d_{AB} + d_{CD} + 2x \tag{5.6}$$

这里 x 是内部分枝的长度。因此，以下两个条件成立：

$$d_{AB} + d_{CD} < d_{AC} + d_{BD} \tag{5.7a}$$

和

$$d_{AB} + d_{CD} < d_{AD} + d_{BC} \tag{5.7b}$$

这两个条件总称为四点条件（four-point condition）。即使可加性只不过近似地存在，这些条件可能也是成立的。

反之，对系统发育关系不知的 4 个 OTU 而言，以上两个条件可用来找出近邻（A 和 B；C 和 D）。一旦这两对近邻被确定，系统树的拓扑图也就被确定了。

萨塔斯和特韦尔斯基（Sattath 和 Tversky，1977）提出以下方法来处理多于 4 个 OTU 的情况。首先，象在 UPGMA 中的情况一样算出距离矩阵。对每种可能的四单位组，比如 OTUi，j，m 和 n，算出 $d_{ij} + d_{mn}$，$d_{im} + d_{jn}$ 和 $d_{in} + d_{jm}$。假定第一个和的值最小，那么，我们把 i 和 j 对以及 m 和 n 对都记 1 分，而 i 和 m 对、i 和 n 对、j 和 m 对、j 和 n 对则都记 0 分。否则，如果 $d_{im} + d_{jn}$ 有最小值，则我们就给 i 和 m 对以及 j 和 n 对记 1 分，而给其他 4 种可能的对子记 0 分。在对所有可能的四单位组都评过分以后，得总分最高的对子即被选为第一个近邻对，并把它当成一个 OTU 来处理，下一步，如在 UPGMA 中的情况一样算出新距离矩阵，然后重复以上过程选出第二个近邻对。此过程继续到所有 OTU 被聚类时止。在我们考虑猿和人的系统发育时将给出此法的详细说明（见第 72 页）。

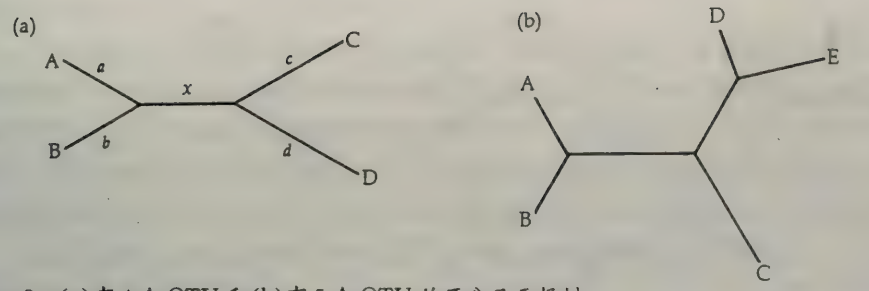


图 5—8 (a)有 4 个 OTU 和(b)有 5 个 OTU 的两分叉无根树。

以近邻关系概念为根据的另一种方法曾由菲奇(Fitch,1981)提出。塞图和根井(Saitou 和 Nei,1987)曾提出一个称为近邻结合法(neighbor-joining method)的方法。这是连续地找出使树的总长度最小的近邻对的方法。

最节省法

最节省或最少进化的原则涉及:找出一个要求最小进化变化数的树,以解释被研究 OTU 间观察到的差异,这样一种树称为最节省树(maximum parsimony tree)。常常可以发现不止一种树有同样的最小变化数,所以,推测出的树可能不是唯一的。

下面讨论的方法最初是为了处理氨基酸顺序数据而提出的(Eck 和 Dayhoff,1966),后来经修改后被用于核苷酸顺序(Fitch,1977)。

我们先对信息位点(informative sites)进行定义。一个核苷酸位点是具有系统发育方面的信息的,仅当它在众多的树中偏向其中一些树时。为了说明信息位点和非信息位点间的区别,让我们考虑以下 4 个假想的顺序:

	位点								
顺序	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	G
3	A	G	A	T	A	T	C	C	A
4	A	G	A	G	A	T	C	C	G
					*		*		*

对 4 个 OTU 有 3 种可能的无根树(图 5—9)。位点 1 不是有信息的,因为所有顺序在该位点上都有 A,以至在 3 种可能的树中都不需要任何变化。在位点 2 上,顺序 1 有 A,而所有其他顺序都有 G,所以只需作一简单假定,即谱系中核苷酸从 G 变成 A 导至了顺序 1。于是,这一位 点也不是有信息的,因为对这 3 种可能的树来说每种都只需要 1 次变化。如图 5—9a 所示,对位点 3 来说,这 3 种可能的树中的每一种都需要 2 次变化,所以也不是有信息的。注意,如果假定图 5—9a 的树 I 中联结 OTU 1 和 2 的节点上的核苷酸是 C(或 A)而不是 G,则该树所需要的变化数仍为 2。图 5—9b 表明对于位点 4 这 3 种树的每一种都要求 3 次变化,所以位点 4 也是非信息的。对于位点 5,树 I 只要求 1 次变化,而树 Ⅱ 和 Ⅲ 则各要求 2 次变化(图 5—9c),所以该位点是有信息的。

从这些例子中我们看到,就我们所谈到的分子数据而言,仅当一个位点上至少有两种不同类型的核苷酸、每种类型至少在两个被研究顺序中出现时,该位点才是有信息的。在上例中,信息位点用星号(*)指出。

为了推出一个最节省树,我们首先要找出所有信息位点。接着,对每种可能的树我们算出每一信息位点上的最小替换数。在上例中,对于位点 5、7 和 9,树 I 分别需要 1、1 和 2 次变化,树 Ⅱ 分别需要 2、2 和 1 次变化,树 Ⅲ 分别需要 2、2 和 2 次变化。最后一个步骤是,对每种树将所有信息位点上的变化数加和,并选出总替换数最小的那种树。在我们的例子中,因为树 I 在信息位点上需要的变化数最

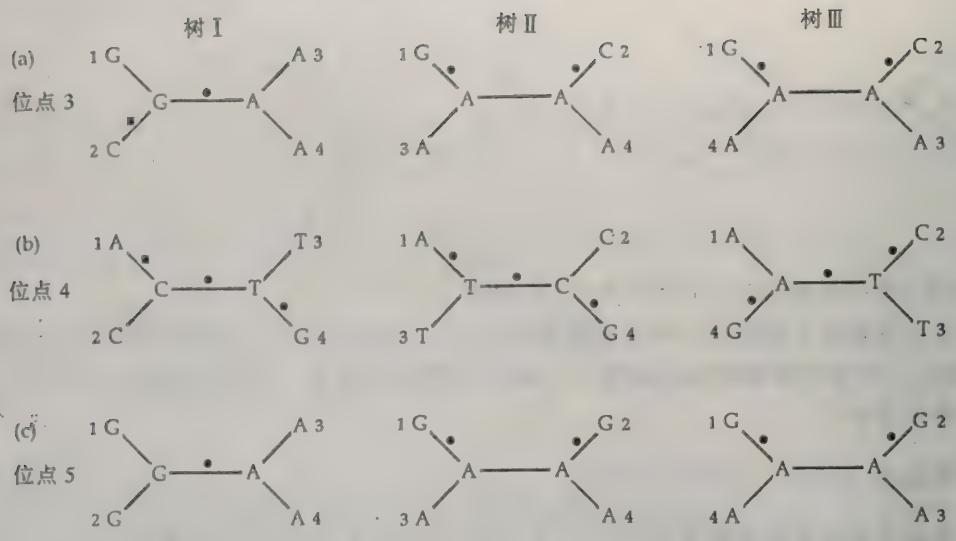


图5－9　关于4个DNA顺序(1、2、3和4)的3种可能的无根树(Ⅰ，Ⅱ和Ⅲ)，这4个顺序曾被用来选取最节省树(见正文)。终端节点处指出了现存物种中该同源位置上的核苷酸种类。分枝上的每一个圆点表示一次替换，它被放在发生替换的分枝上。注意，每种树的两个内部节点上的核苷酸只表示了几种可供选择的构建中的一种。例如，树Ⅲ(c)的两个内部节点上可能都是G而不是A(右下角)。在这种情况下，两次替换将位于导致物种3和4的分枝上。而所要求的替换最小数却保持相同。

小(4)，所以选中树Ⅰ。

在有4个OTU的情况下，一个信息位点只偏向3种可供选取的树中的一种。例如，位点5偏向树Ⅰ甚过树Ⅱ和树Ⅲ，因而被说成是支持树Ⅰ的。容易看出，受信息位点支持数最大的那种树就是最节省树。例如，上例中，树Ⅰ受2个位点支持，树Ⅱ受1个位点支持，而树Ⅲ无位点支持，结论不言自明。

5.4　表型学与进化枝学

分类学中长期存在的争论就是那些一直在“进化枝学家”和“表型学家”间发生的言辞刻薄的争论。进化枝学(cladistics)这个词可以定义为研究进化途径的科学。换句话说，进化枝学家对这样一些问题感兴趣：在一群生物中存在多少分枝？哪些分枝与别的哪些分枝相联？分枝次序如何(Sneath和Sokal，1973)？一种表示这种祖先—后代关系的树状网络称进化分枝图(cladogram)。换句话说，进化分枝图即指一种有根系统树的拓扑图。

另一方面，表型学(phenetics)是根据一群生物间的相似程度来研究它们间的关系的科学，引以为据的类似性可以是分子上的、表型上的或解剖学上的。表示表型学关系的树状网络称表型关系图(phenogram)。虽然表型关系图可以当作进化枝学关系的指示物来用，但它可不必与进化分枝图等同。如果在分岐时间和遗传学上的(或形态学上的)分岐程度间存在线性关系，则这两种树可能会变得相互等同。

在上面所讨论的方法中，最节省法是进化枝学方法的一个典型代表，而UPGMA法则是一种典型的表型学方法。然而，其他方法却不能按以上标准简单地进行分类。例如，变形距离法和近邻关系法常常被说成是表型学方法，但这是一种不确切的说法。虽然这些方法用了类似性(或不似性，即距离)尺度，但它们并没有假定类似性和进化关系间直接关联，也并不打算推测表型学关系。

在分子系统发育中，用更合理一些的分类方法应把它们分成距离法和特性状态法(distance and character-state approaches)。属于前者的方法根据距离尺度，象核苷酸或氨基酸的替换数等，而后者则依靠该特性的状态，如某一特定位点上的核苷酸或氨基酸，或某一DNA位置上缺失或插入的出现或

不出现等。根据这种分类，则 UPGMA 法、变形距离法和近邻关系法都是距离法，而最节省法则是一种特性状态法。

一直有这样一种说法，即，特性状态法比距离法更有效用，因为该原始资料是一串特性状态（例如核苷酸顺序），而将特性状态数据转变成距离矩阵以后，有些信息就丢失了。然而我们注意到，虽然最节省法事实上应用的是原始资料，但通常它仅用了可用资料的一小部分。例如，第 75 页上的例子中，只有 3 个位点被用于分析中，倒有 6 个位点被排除在外了。为此，此法常比有些距离矩阵法的效率低（例如见 Saitou 和 Nei，1986）。当然，如果信息位点数多，则最节省法一般是非常有效的。

最后，应该注意到，不管用何种方法，一个推测树常难免有拓扑图错误。为了得到正确的树，通常需要大量数据资料。

5.5 枝长的估计

除了在 UPGMA 法中外，我们都不曾讨论过怎样去估计枝长的问题。现在，我们在该树的拓扑图已被推测出来的假定下，来处理这一问题。我们只考虑用距离矩阵法推测出来的树对于最节省法来说，这个问题要复杂得多（见 Fitch，1971）。下面讨论的方法是菲奇和马戈利阿什（Fitch 和 Margoliash，1967）的方法。

首先，我们考虑一种最简单的情形，即一个有 3 个 OTU（A，B 和 C）和 1 个节点的无根树（图 5－10a）。设 x、y 和 z 分别是导向 A、B 和 C 的分枝的长度。很容易看出以下等式成立：

$$d_{AB} = x + y \tag{5.8a}$$

$$d_{AC} = x + z \tag{5.8b}$$

$$d_{BC} = y + z \tag{5.8c}$$

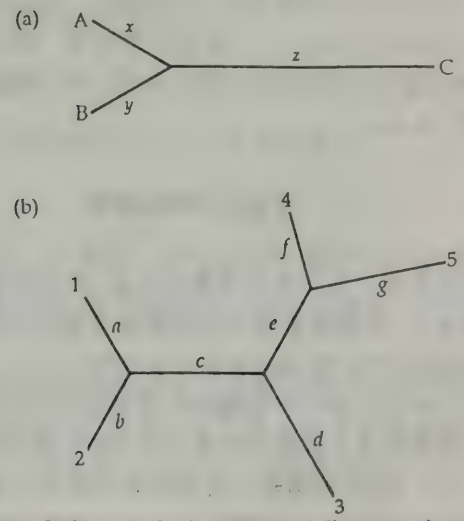


图 5－10　用菲奇和马戈利阿什（Fitch 和 Margoliash，1967）法计算枝长时采用的无根系统树。(a)有 3 个 OTU 的树，(b)有 5 个 OTU 的树。

从这些等式，我们得到以下解：

$$x = \frac{d_{AB} + d_{AC} - d_{BC}}{2} \tag{5.9a}$$

$$y = \frac{d_{AB} + d_{BC} - d_{AC}}{2} \tag{5.9b}$$

$$z = \frac{d_{AC} + d_{BC} - d_{AB}}{2} \tag{5.9c}$$

我们现在来处理有 3 个以上 OTU 的情形。为简单计，我们假定有 5 个 OTU（1，2，3，4 和 5），而其拓扑图和各枝长如图 5－10b 所示。假定 OTU 1 和 2 是构树过程中最先被聚类的 OTU，则我们分别用 A 和 B 来表示 OTU1 和 2，而把所有别的 OTU（3、4 和 5）放在一个复合 OTU 中，并表示成 C。通过

这种安排，我们可以用等式 5.9a－5.9c 来估计导向 A、B 和 C 的分枝的长度。不过，现在 $d_{AC}=d_{1(345)}=(d_{13}+d_{14}+d_{15})/3$，$d_{BC}=d_{2(345)}=(d_{23}+d_{24}+d_{25})/3$ 于是，我们有 a＝x 和 b＝y。接着 OTU1 和 2 被看成一个复合 OTU。下面再假定，复合 OTU(12)与单 OTU3 是下一个要结合在一起的对子。于是，我们分别把复合 OTU(12)和单 OTU3 用 A 和 B 表示，把其余的 OTU(即 4 和 5)放在新的复合 OTUC 中。以上述同样的方式，我们将得到 x，y 和 z。注意，这时 d＝y，c＋(a＋b)/2＝x。从前一次我们得到的关于 a 和 b 的值，可以算出 c。继续进行这样的过程，直到所有枝长都得到时为止。

注意，有时某一被估出的枝长可能会是负的。由于实际枝长绝不可能是负的，所以，最好用 0 来代替这样的估值。

作为应用上述方法的一个例子，让我们来算一下图 5－7c 中那个树的各枝长。为便于说明，我们把在推测该树的拓扑图时用过的距离矩阵再次列出。为了避免与等式 5.9 中的符号混淆，我们把 OTUA，B，C 和 D 分别地重新命名为 OTU1，2，3 和 4。

	OTU		
OTU	1	2	3
2	8		
3	7	9	
4	12	14	11

因为 OTU1 和 2 最先被聚类，所以，我们把 OTU3 和 4 放在一个复合 OTUC 中，从而先算出导向 OTU1 和 2 的两个分枝的长度(a 和 b)。而我们有 $d_{AB}=d_{12}=8$，$d_{AC}=(d_{13}+d_{14})/2=(7+12)/2=9.5$ 和 $d_{BC}=(d_{23}+d_{24})/2=11.5$。由等式 5.9a－5.9c，我们有：a＝x＝(8＋9.5－11.5)/2＝3，b＝y＝(8＋11.5－9)/2＝5。下一步我们把 OTU1 和 2 处理成 OTU(12)并用 A 表示。于是，我们有：$d_{AB}=d_{(12)3}=(d_{13}+d_{23})/2=(7+9)/2=8$，$d_{AC}=d_{(12)4}=(d_{14}+d_{24})/2=(12+14)/2=13$，$d_{BC}=d_{34}=11$。因此，根据等式 5.9a－5.9c，我们有：x＝(8＋13－11)/2＝5，d＝y＝(8＋11－13)/2＝3，e＝z＝(13＋11－8)/2＝8。由图 5－7c 可以看出(a＋b)/2＋c＝x，所以，c＝1。至此计算完毕。然而请注意，由于我们不知道根的准确位置，所以我们不能估出联系根和 OTUD 的分枝的长度，而只能估出从 OTUA，B 和 C 的共同祖先节点通过根到 OTUD 的长度，即 e＝8。

5.6 寻找无根树的根

大多数构树法得到的都是无根树。为了找到无根树的根，我们通常都需要一个组外单位(即一个 OTU，其外在信息，如古生物学证据，清楚地表明它比那些被研究的分类单位更早地发生分枝)。而根则被放在该组外单位和将它和别的 OTU 联结起来的节点之间。

虽然我们必须能肯定该组外单位的确比所有别的分类单位更早地发生分岐，但奉劝各位，不要选一个与被研究的各分类单位亲缘关系过于遥远的组外单位，因为，在这样的情况下很难得到该组外单位与其他分类单位间的可靠估值。例如，在构建一组有胎盘哺乳类间的系统发育关系时，我们可以用一种有袋类作为组外单位。仅当所用的 DNA 序列在进化中是高度保守的时，鸟类才可能被当成可靠的组外单位。在此例中植物与真菌显然也够资格作为组外单位，不过，仅仅因为它们与哺乳类的关系太过遥远用它们作组外单位，结果就可能会出现严重的拓扑图错误。用一个以上的组外单位(假若它们与别的分类单位都不是关系遥远)，则一般能改善该树拓扑图的估计结果。该组外单位还必须在系统发育上与别的 OTU 不至于太接近，因为，在这种情况下我们不能肯定它是一个真正的组外单位，即不能肯定它与其他 OTU 的分岐早于这些 OTU 相互间的分岐。

在缺少组外单位的情况下，我们可以通过假定进化速率在所有分枝上近似一致来放置根。在这种假定下，我们把根放在两个 OTU 间的最长路线的中点处。

5.7 物种分岐时间的估计

因为古生物学记录远不够完全，所以我们常常不去管物种分岐的年代。DNA 顺序数据在这方面

可能会给出很大帮助。从以前的研究中获知我们假定某一 DNA 序列的进化速率，为每年每位点 r 替换。为了得到物种 A 和 B 间的分歧时间(T)，我们对来自这两个物种的该顺序进行比较，并算出每位点替换数(K)。如第四章所示，我们有以下等式：

$$r = \frac{K}{2T} \tag{5.10}$$

因此，T 估计为

$$T = \frac{K}{2r} \tag{5.11}$$

如在第四章所说明过的那样，从一组生物得到的核苷酸替换速率也许不能用于另一组生物。为了避免出现这样的问题，我们可通过加进第 3 个物种 C(它与物种对 A 和 B 的分歧时间已知)，来估计替换速率(图 5—11)。设 Kij 是物种 i 和 j 间每位点核苷酸替换数，那么，核苷酸替换速率可由：

$$r = \frac{K_{AC} + K_{BC}}{2(2T_1)} \tag{5.12}$$

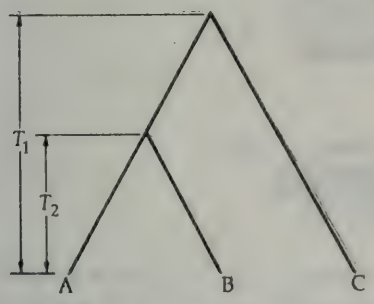


图 5—11　估计分歧时间用的模型树(见正文)

估出。物种 A 和 B 间的未知分歧时间(T_2)用：

$$T_2 = \frac{K_{AB}}{2r} = \frac{2K_{AB}T_1}{K_{AC} + K_{BC}} \tag{5.13}$$

估出。反之，在 T_2 已知而 T_1 未知的情况下，T_1 由下式给出：

$$T_1 = \frac{(K_{AC} + K_{BC})T_2}{2K_{AB}} \tag{5.14}$$

以上公式均假定速率恒定。但如前一章中所讨论的那样，此假定常常是不成立的，所以估出的分歧时间应该小心处理。李和谷村(Li 和 Tanimura，1987)曾提出一种方法，此法能降低替换速率不等对分歧时间估计的影响。

早先我们已经注意到，两顺序间的分歧时间可能比携带这两个顺序的两物种间的分歧时间早。不过，如果我们关心的是长期分歧事件，比如说几百万年或更长的时间等级，则这种误差通常并不很严重。还应该注意，分歧时间的估值通常承受着较大的随机误差。为了减少这类误差，在估计中应该用许多顺序。

5.8　进化枝

系统发育研究的目的是建立不同物种间的进化关系。特别地，我们对找出自然进化枝(clades)感兴趣。一个进化枝被定义成一组有共同祖先的物种，而该祖先却不为这一进化枝外的其他物种所共有。

图 5—12 显示脊椎动物的 3 个纲：鸟纲、爬行纲和哺乳纲间的进化关系。我们看到，经典分类学把爬行类指定为一个独立的纲，这与进化枝的定义不符，因为这 3 类爬行动物与另一个类群——鸟类有共同祖先，而后者则并不被包括在爬行纲的定义范围内。另一方面，鸟类与鳄类倒的的确确地构成了一个自然进化枝，初龙亚纲，因为它们所共有的一个共同祖先不为除鸟类和鳄类以外的任何现存生物所共有。类似地，所有鸟类和所有爬行类合在一起构成一个自然进化枝。

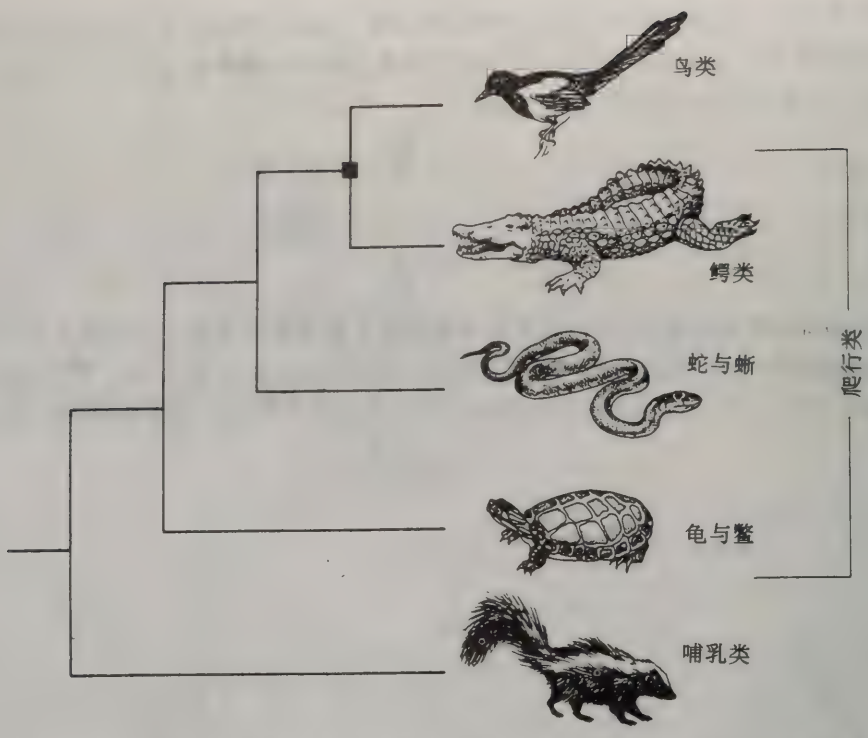


图 5－12　鸟类，爬行类和哺乳类的分枝进化图。爬行类不构成一个自然进化枝，因为它们与鸟类有共同祖先，而后者并不被包括在爬行纲中。另一方面，鸟类与鳄类却构成一个自然进化枝（初龙亚纲），因为它们所共有的一个共同祖先（黑色方块）不为任何别的生物所共有。

5.9　人和猿的系统发育

与人类在进化上最接近的亲缘关系问题一直是令生物学家们感兴趣的问题。例如，达尔文（Darwin，1871）曾认为，非洲猿，黑猩猩（*Pan*）和大猩猩（*Gorilla*）是人类的近亲，所以他的结论是，人类的进化起源将被发现一定是在非洲。。由于各种原因达尔文的观点没有引起人们的兴趣，且在一个相当长的时期里，分类学家们认为人属（*Homo*）是唯一与猿的亲缘关系较远的一类，所以它自成一科，即人科。黑猩猩、大猩猩和马来猩猩（*Pongo*）通常被放在独立的一个科，即猩猩科（图 5－13a）中。长臂猿（*Hylobates*）或者独立成科或者被分在猩猩科（图 5－13b；见 Simpson，1961）。古德曼（Goodman，1963）曾正确地认识到，这种系统安排是以人类为中心的，因为人类代表着“系统发育所达到的一个新等级，一个比猿和所有别的以前存在的等级更高的等级”。然而，把各种猿放在一个科里而把人放在另一个科里，这就意味着这些猿相互间共有一个比人更近的共同祖先。当把人与一种现存的猿放在同一进化枝中时，它通常是与亚洲猿，即马来猩猩放在一起（图 5－13c；Schultz，1963）。

由于血清沉淀法的应用，古德曼（Goodman，1962）已能证明人、黑猩猩和大猩猩构成一个自然进化枝（图 5－13d），而马来猩猩和长臂猿与别的猿类发生分岐的年代则要早得多。根据微量互补结合资料，萨里奇和威尔逊（Sarich 和 Wilson，1967）估计人与黑猩猩或大猩猩间的分岐时间可能近至 500 万年前，而不是当时普遍为古生物学家所接受的至少 1500 万年前。

然而，血清学方法、电泳法以及氨基酸顺序等都不能解决人与非洲猿间的进化关系问题，而所谓人—大猩猩－黑猩猩的三分法仍未得到解决，并继续是一个极有争议的问题（图 5－14）。下面，我们将用 M. 古德曼及其同事（见 Miyamoto 等，1987）和其它学者（Maeda 等，1988）得到的 DNA 顺序资料，以表明分子证据支持人－黑猩猩进化枝，同时对前几节所讨论的构树法加以解释说明。

表 5－2 展示了以下几个 OTU 每一对子间每 100 位点的核苷酸替换数：人（H），黑猩猩（C），大猩猩（G），马来猩猩（O）和罗猴（R）。我们先对这些距离用 UPGMA 法。人与黑猩猩间的距离最短（$d_{HC}=1.45$），因此，我们首先把这两个 OTU 结合，并把节点放在 $1.45/2=0.73$ 处（图 5－15a）。然后，我们

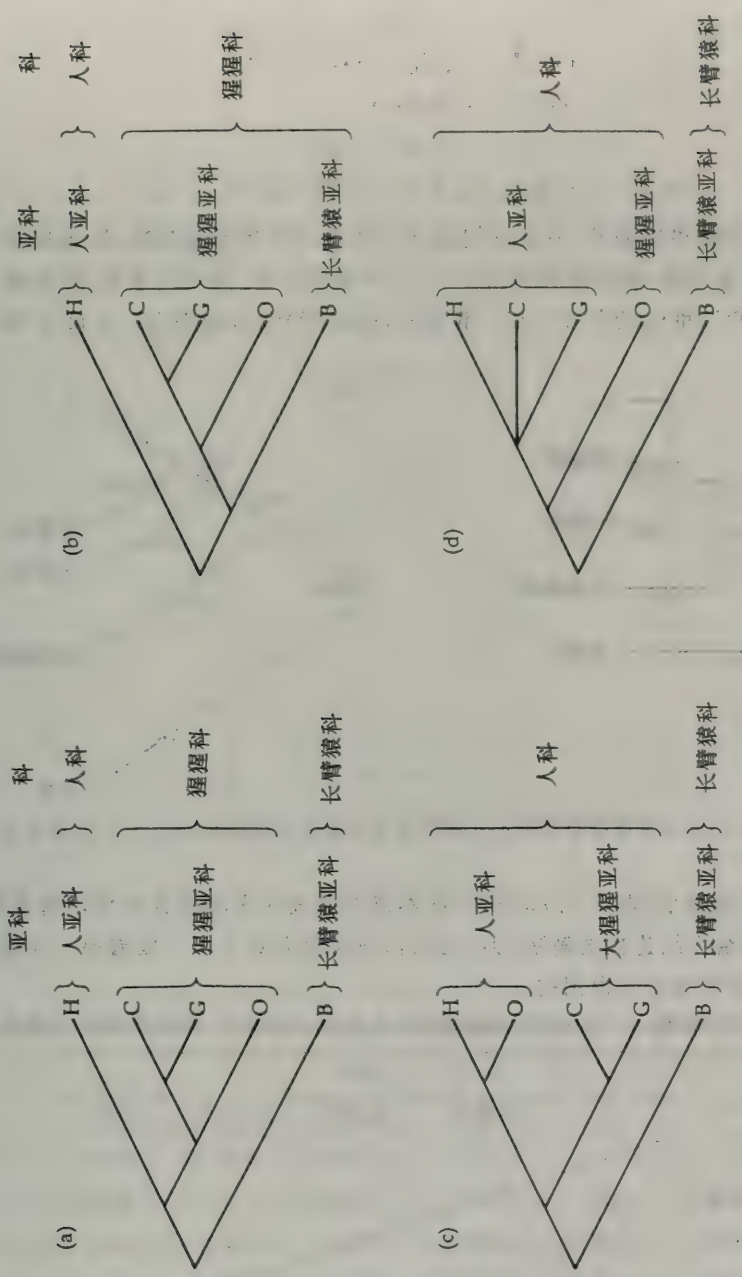


图 5—13　现代猿和人(人总科)的 4 种供选择的系统发育和分类方式。传统分类法把人类独立出来，如(a)和(b)中所展示的那样。人与马来猩猩的聚类如(c)中所示。把分子的和形态学的证据结合起来则偏向(d)中所示的那种分类方式。

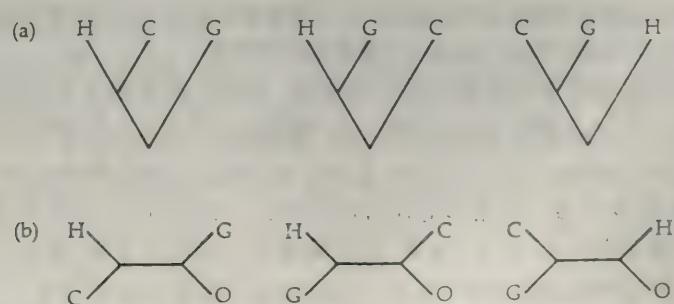


图 5—14　(a)关于大猩猩、黑猩猩和人的 3 种可能的有根树。(b)以马来猩猩作为组外单位的相应无根树。物种缩写：C，黑猩猩(*Pan troglodytes*)；H，人(*Homo sapiens*)；G，大猩猩(*Gorilla gorilla*)；O，马来猩猩(*Pongo pygmaeus*)

算出该复合OTU(HC)与其余每一物种间的距离，得出一个新的距离矩阵：

	OTU		
OTU	(HC)	G	O
G	1.54		
O	2.96	3.04	
R	7.53	7.39	7.10

因为(HC)和G为最短的距离所隔开，所以它们是下一个要结合起来的对象，且该联结的结节位于1.54/2=0.77处。继续此类过程，我们即得到图5－15a中那样的树。我们注意到，估出的关于H和C的分枝结点非常接近于(HC)和G的分枝节点。事实上，这两个节点间的距离，比H、C和G间所有关于

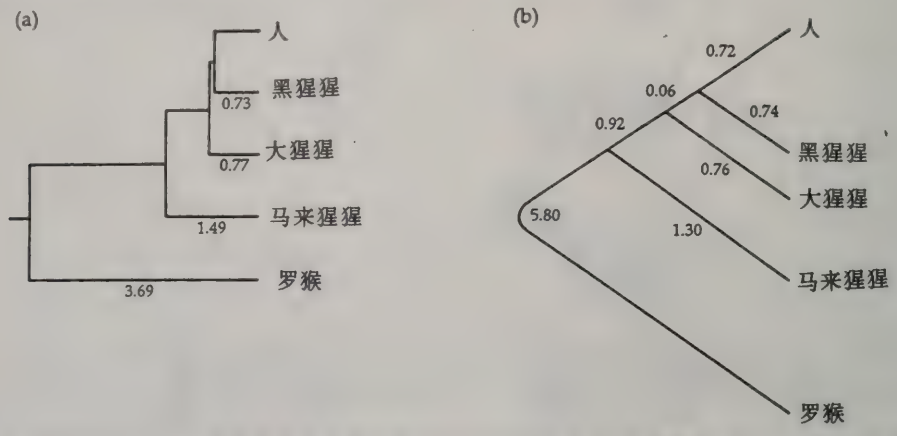


图5－15　由UPGMA法(a)以及萨塔特和斯韦尔斯基的近邻关系法(b)推测出的、关于人、黑猩猩、大猩猩和罗猴的进化树。

成对距离估值的标准误差都小(表5－2)。所以，虽然该资料表示人的最近现存亲属是黑猩猩，但该资料并没对分枝次序问题给出结论性的解决。另一方面，马来猩猩对于人—黑猩猩—大猩猩进化枝而言是一个组外单位，这一点则是毫不含糊的。

表5－2　各OTU间每100位点核苷酸替换数的平均值(对角线下)和标准误差(对角线上)[a]

	OTU				
OTU	人	黑猩猩	大猩猩	马来猩猩	罗猴
人		0.17	0.18	0.25	0.41
黑猩猩	1.45		0.18	0.25	0.42
大猩猩	1.51	1.57		0.26	0.41
马来猩猩	2.98	2.94	3.04		0.40
罗猴	7.51	7.55	7.39	7.10	

自Li等(1987b)

a. 所用顺序资料为长5.3Kb的非编码DNA，它由两个分开的区域构成：(1)由Koop等(1986b)报导的η-珠蛋白基因座位(2.2kb)和由Maeda等(1983,1988)定序的长3.1kb的η－δ珠蛋白基因间区域。

下面，我们用萨塔特和斯韦尔斯基的近邻关系法。我们一次考虑4个OTU，由于现有5个OTU，所以有5!/〔4!(5－4)!〕＝5种可能的四单位组。我们从OTUH、C、G和O开始，并算出以下距离之和(数据自表5－2)：$d_{HC}+d_{GO}=1.45+3.04=4.49$，$d_{HG}+d_{CO}=4.45$和$d_{HO}+d_{CG}=4.55$。因为第2个和最小，所以我们选H和G为一个近邻对，C和O为另一对(表5－3)。类似地，我们再考虑另外4种可能的四单位组，结果如表5－3所示。从表5－3的最后一行我们看到，其在所有近邻对中(OR)有最高的近邻关系得分，所以我们选取(OR)为第一个近邻对。把这个对子当作一个OTU，则我们得到以下新的距离矩阵：

	OTU		
OTU	H	C	G
C	1.45		
G	1.51	1.57	
(OR)	5.25	5.25	5.22

表 5—3　用表 5—2 中的距离矩阵得到的近邻关系得分

用于比较的 OTU[a]	成对距离之和	选取的近邻对
H,C,G,O	$d_{HC}+d_{GO}=4.49$ $d_{HG}+d_{CO}=4.45$ $d_{HO}+d_{CG}=4.55$	(HG),(CO)
H,C,G,R	$d_{HC}+d_{GR}=8.84$ $d_{HG}+d_{CR}=9.06$ $d_{HR}+d_{CG}=9.08$	(HC),(GR)
H,C,O,R	$d_{HC}+d_{OR}=8.55$ $d_{HO}+d_{CR}=10.53$ $d_{HR}+d_{CO}=10.45$	(HC),(OR)
H,G,C,R	$d_{HG}+d_{OR}=8.61$ $d_{HO}+d_{GR}=10.37$ $d_{HR}+d_{GO}=10.55$	(HG),(OR)
C,G,O,R	$d_{CG}+d_{OR}=8.67$ $d_{CO}+d_{GR}=10.33$ $d_{CR}+d_{GO}=11.59$	(CG),(OR)
总分	(HC)=2,(HG)=2,(HO)=0,(HR)=0,(CG)=1 (CO)=1,(CR)=0,(GO)=0,(GR)=1,(OR)=3	

a. H,人;C,黑猩猩;G,大猩猩;O,马来猩猩;R,罗猴。

由于仅留下 4 个 OTU,所以很容易看出,$d_{HC}+d_{G(OR)}=6.76<d_{HG}+d_{C(OR)}=6.76<d_{H(OR)}+d_{CG}=6.82$。因此,我们选 H 和 C 作为一个近邻对,G 和(OR)为另一近邻对。用此法最后得到的树如图 5—15b 所示。该树的拓扑图与图 5—15a 中的图等同。然而请注意,在此法中最先相互结合成对的是 O 和 R,而不是 H 和 C。这是因为,在无根树中 O 和 R 事实上就是近邻。图 5—15b 中的各枝长由更早些时给出的方法估出。

最后,我们看看最节省法。为简单起见,我们只考虑人、黑猩猩、大猩猩和马来猩猩。表 5—4 列出了包括 η—珠蛋白假基因及其周围区域的、长 10.2kb 的片段中,有关的信息位点(Koop 等,1986a; Miyamoto 等,1987; Maeda 等,1988)。每一位点所支持的假定在最后一列中给出。如果我们只考虑碱基改变,则 15 个信息位点中,有 8 个支持人—黑猩猩进化枝(假定Ⅰ),4 个支持黑猩猩—大猩猩进化枝(假定Ⅱ),3 个支持人—大猩猩进化枝(假定Ⅲ)。而且,4 个涉及裂缝的信息位点全部支持人—黑猩猩进化枝。因此,人—黑猩猩进化枝被选作最可能是真实系统发育方式的代表。在一次更详细的分析中,威廉姆斯和古德曼(Williams 和 Goodman,1989)曾证明,有关支持人—黑猩猩进化枝的结果,有按 1%水平衡量的统计意义。

表 5—4　人、黑猩猩、大猩猩和马来猩猩的顺序中的信息位点

位点[a]	顺序 人	黑猩猩	大猩猩	马来猩猩	受支持的假定[b]
来自 Miyamoto 等(1987)的数据					
34	A	G	A	G	Ⅱ
560	C	C	A	A	Ⅰ
1287	*[c]	*	T	T	Ⅰ
1338	G	G	A	A	Ⅰ
3057—3060	* * * *	* * * *	TAAT	TAAT	Ⅰ
3272	T	T	*	*	Ⅰ
4473	C	C	T	T	Ⅰ
5153	A	C	C	A	Ⅱ
5156	A	G	G	A	Ⅱ
5480	G	G	T	T	Ⅰ
6368	C	T	C	T	Ⅱ
6808	C	T	T	C	Ⅱ
6971	G	G	T	T	Ⅰ
来自 Maeda 等(1988)的数据					
127—132	* * * * * *	* * * * * *	AATATA	AATATA	Ⅰ
1472	G	G	A	A	Ⅰ
2131	A	A	G	G	Ⅰ
2224	A	G	A	G	Ⅱ
2341	G	C	G	C	Ⅱ
2635	G	G	A	A	Ⅰ

自 Williams 和 Goodman(1989)修改而成。

a、位点序数对应于原始来源中给出的序数。所用序列的总长度为 10.2 kb，约为表 5—2 中所用序列的 2 倍。

b、假定：Ⅰ，人和黑猩猩在一个进化枝中；Ⅱ，黑猩猩和大猩猩在一个进化枝中；Ⅲ人和大猩猩在一个进化枝中。

c、一个星号表示在该位点上缺失一个核苷酸。

另外还有些类型的分子数据也支持人和黑猩猩聚类的结论。例如，DNA—DNA 杂交数据(Sibley 和 Ahlguist，1984；Caccone 和 Powell，1989)明确地表示，在图 5—14 里 3 种供选取的系统发育关系中，人和黑猩猩聚被类成一个进化枝。于是，与人关系最近的现存亲属是两种黑猩猩，其后为大猩猩、马来猩猩和 9 种长臂猿。

5.10　线粒体和叶绿体的内共生起源

基本上有两种理论解释真核生物中与细胞核分开的线粒体和叶绿体基因组的存在。第一种理论(例如，Cavalier-Smith，1975)认为，细胞器的基因是自体起源的，是通过直接的父子关系从细胞核基因传下来的。细胞核基因组的某些部分被并入到一个由膜包被的细胞器中，接着假定它的准独立存在，从而完成了这一起源过程。与其不同的是，内共生学说(例如 Margulis，1981)认为核外 DNA 的起源是外源性的。按照这种看法真核生物的祖先吞噬了原核生物，接着，由于互利或有共生关系的缘故，后者被保留了下来。随着时间的流逝，这种内共生通过某些基因的丢失而变得更为密切，以至最后成了强制性共生者(即，不能在其宿主体外独立地存在的生物)。

现在，分子证据绝对地支持内共生学说。使叶绿体和原核生物的基因组与真核生物的基因组区分开来的生化特性，列于表 5—5 中。然而，最根本的支持来自 rRNA 顺序数据。由于 rRNA 顺序有较低的替换速率，所以，它们已被证明对处理涉及非常古老的进化分歧事件的问题是很有用的。

施瓦茨和科塞尔(Schwarz 和 Kossel，1980)证明，来自玉米(*Zea mays*)叶绿体中的 16SrRNA，其核苷酸顺序与来自大肠杆菌(*E. coli*)的 16SrRNA 的极为相似。在细胞核的和叶绿体的 rRNA 顺序间，这种类似程度就要低得多。来自光合蓝细菌的 16SrRNA 顺序的详细分析(Giovannovi 等，1988)支持这样的看法：即，绿色的叶绿体来自一群称为蓝细菌的光合细菌(Bonen 等，1979)(见图 5—16a)。

表 5－5　将叶绿体和原核生物的基因组与真核生物的核基因组区别开来的分子特征

1. 无组蛋白 DNA
2. 大小为 120,000－150,000 碱基对
3. 环状基因组
4. 转录的利福平敏感性
5. 核糖体受链霉素、氯霉素、spectromycin 和巴龙霉素抑制
6. 翻译对放线菌酮的不敏感性
7. 翻译以甲酰甲硫氨酸为起始
8. mRNA 的多聚腺苷缺失或很短
9. 原核促进子结构

rRNA 顺序的系统发育分析表明，线粒体来自紫色细菌的 α 分枝（图 5－16a；Cedergren 等，1988）。然而，所用的细胞核 rRNA 顺序指出，高等植物在大约与动物和真菌相同的时刻发生分枝（图 5－16b），这与传统观点一致；而线粒体 rRNA 顺序却指出，高等植物在非常接近紫色细菌的根处聚类，与真菌、绿色藻类和动物则是分离的（图 5－16a）。这是与传统观点，即把高等植物和绿色藻类组合在一个进化枝中的观点相矛盾的。为此缘故，格雷等（Gray 等，1989）曾提出，高等植物的线粒体中的 rRNA 基因，有比其他生物线粒体中的 rRNA 基因更近的进化起源。此假说是否会被进一步的证据所支持，我们将拭目以待。

5.11　分子古生物学

从微克量保存下来的组织中得到的不纯样本的 DNA，现在已有可能对它的片段加以定序了。所采用的方法是聚合酶链式反应法（PCR）。PCR 是通过应用两种已知引物来使某序列混合物中的一种独特序列扩增的方法（图 5－17；Saiki 等，1985，1988；Scharf 等，1986；Engelke 等，1988）。每一引物和一个 DNA 的互补小段附着，从而引发 DNA 多聚酶的结合，然后拷贝该片段。由于每一新产生的拷贝都可充作进一步复制时的模板，所以靶片段的拷贝数将呈指数增长。用此法有可能将混合物中的某一被选定的 DNA 片段合成产生许多拷贝，此时混合物中其他 DNA 序列可能会远为超量地存在（Kocher 等，1989）。因此，采用 PCR 我们可以从博物馆标本，象保存下来的有机物质（主要是皮肤和肌肉）、严重损坏的考古学遗物甚至骨骼中找回某些特别的 DNA 序列（Hagelberg 等，1989）。

应用此法现在已可能建立某些已灭绝物种，象南非斑驴和澳大利亚袋狼间的系统发育关系，或决定斯堪的那维亚铁器时代的沼泽地人和埃及木乃伊，这些已灭绝人类群体间的祖先-后裔关系。对后两个人类群体的形态学比较，得到的结果比较含糊。

应用 PCR 法，托马斯等（Thomas 等，1989）已能对来自袋狼 *Thylacinus cynocephalus* 和其他澳大利亚与南美有袋类的线粒体的 219 个核苷酸加以定序和比较，并且与来自有胎盘类的同源顺序比较。应用此法，他们已能在以下两种说法间作出取舍：(a)袋狼与南美的有袋动物类群有亲缘关系，和(b)袋狼与其他澳大利亚有袋类的亲缘关系很近。从这些顺序比较中得出的结论是，袋狼与另外两种澳大利亚有袋类，几近灭绝的塔斯马尼亚魔鬼（一种袋獾，*Sarcophilus harrisii*）和澳大利亚虎猫（一种袋鼬，*Dasyurus maculatus*）有很近的亲缘关系，但与一种南美有袋类，负鼠（一种毛鼻，*Philander opossum andersoni*）只有较疏远的亲缘关系（图 5－18）。于是，袋狼和南美有袋动物间形态学上的相似，被认为是一个形态学水平上趋同进化的例子，它在线粒体 DNA 中没有平行关系。

5.12　深色海滩雀：物种保护生物学中的一次教训

最后一只深色海滩雀死于 1987 年 6 月 16 日，地点在佛罗里达州俄伦多附近沃特．迪斯尼世界的动物园里。深色海滩雀在 1872 年被发现。因其黑色的外貌而被明确地归类成一个亚种（*Ammodramus maritimus nigrescens*）。*A. m. nigrescens* 的地理分布被限制在佛罗里达州布列伐得县中的盐碱沼泽

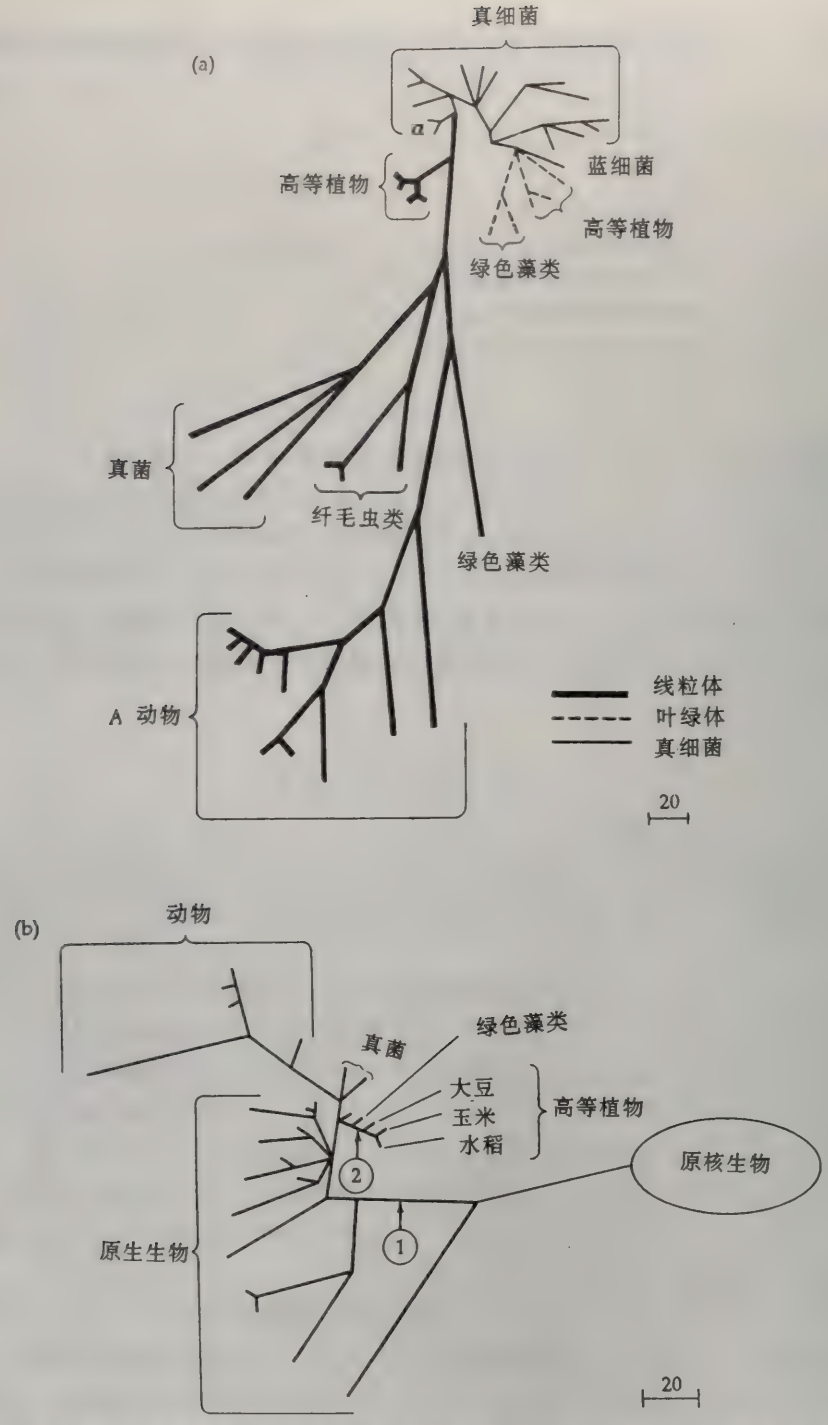


图 5－16　(a)从小亚基 rRNA 的顺序推测出的无根树中真细菌　叶绿体　线粒体部分。注意，来自绿色藻类和植物的叶绿体顺序都是单源发生的，而线粒体顺序则是多源发生的。(b)从细胞核小亚基 rRNA 顺序推测出的无根树。步聚 1(用圆圈标出)是早期内共生、假定已产生了大多数真核生物的线粒体基因组，步聚 2 是后期共生，Gray 等(1989)假定它贡献了高等植物线粒体的 rRNA 基因。每一分枝的长度与其两个内部点间的替换数成比例，比例尺度如图中所示。自 Gray 等(1989)修改而成。

地带(图 5－19)，到了 1980 年，自然界中只能找到 6 个个体，全是雄性。显然该群体的灭绝已命中注定，于是，一个人工杂交计划被当作为保护该亚种的基因的最后一搏而展开了。

在这种情况下，保护方案就是将这些濒临灭绝的亚种的雄性个体，与来自我们所能找到的与其亲缘关系最近的亚种的雌体交配。然后，子一代的杂种雌体再与该雄体回交，它们产生的后代又再次与原雄体回交，此过程循环往复直至该雄体死去为止。该实验的关键是要决定应从哪种群体中选取雌体，即决定哪一个亚种在系统发育上最接近该濒临灭绝的亚种。

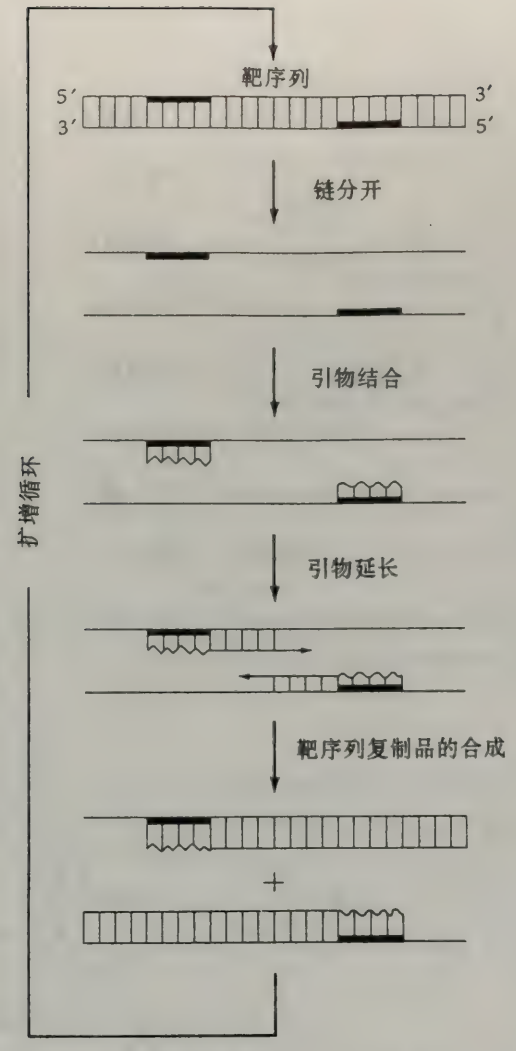


图 5—17　用聚合酶链式反应法(PCR)使混合物中的特定 DNA 序列扩增。DNA 经加热解链成单链分子。两个合成的 DNA 小片段(锯齿状波纹线)各与靶序列(黑色粗线)一端上的特别顺序互补,充作引物。每一引物各与不同链上的互补顺序结合。然后,DNA 多聚酶通过在引物上添加核苷酸来使其延长。很快地,靶序列的精确复制品即产生了。在以后的循环中,原始靶序列及其复制品都可充作模板。相反,不含与引物互补的顺序的 DNA 分子则不被扩增。因此,不纯的 DNA 混合物也可用来只对其中的一种序列进行扩增,即使在其他序列远为过量地存在的情况下,这一点也能办到。更详细说明见 Mullis(1990)。

在海滩雀的例子中,有 8 个可区别的亚种供选取。这些物种的地理分布范围如图 5—19 所示。根据形态学和行为特征所作出的决定是,最接近深色海滩雀的亚种为斯科特的海滩雀(*A. m. peninsulae*),它们栖息在佛罗里达湾的沿岸。在此决定下,*nigrescens* 亚种的几个雄体与来自 *peninsulae* 亚种的雌体交配。已取得了两次成功的回交,而由此产生的群体则此后维持近交,以期有一天能将这"重建"的亚种放回到它原来的栖息地中去。

为了看一看该雌体的选择是否正确,阿维斯和纳尔逊(Avise 和 Nelson,1989)曾进行过有关线粒体 DNA 的限制酶模式的比较,材料的一方来自已故的纯种 *nigrescens* 的标本,另一方来自海滩雀现存 8 个亚种中的 5 个,共 39 个个体。他们选择线粒体 DNA 有几个原因。首先,脊椎动物中线粒体 DNA 已知进化非常迅速(第四章),因此它可以为区别亲缘关系较近的生物提供一个分辨率较高的解决途径。其次,线粒体是母系遗传的,所以不会发生因等位基因分离而造成的复杂局面。最后,由于母系传递模式来自已故雄性深色海滩雀中的线粒体 DNA,并不曾转移到复原育种方案中的杂种个体里。所以,虽然细胞核中的某些基因已在杂种中生存了下来,但深色海滩雀的线粒体基因却不一样,它们是真正的灭绝了。

根据限制酶模式,阿维斯和纳尔逊(Avise 和 Nelson,1989) 用 UPGMA 法和最节省法,构建出了海滩雀几个亚种间的进化关系(图 5—20)。从图中可以看到,大西洋沿岸群体,包括深色海滩雀在内,

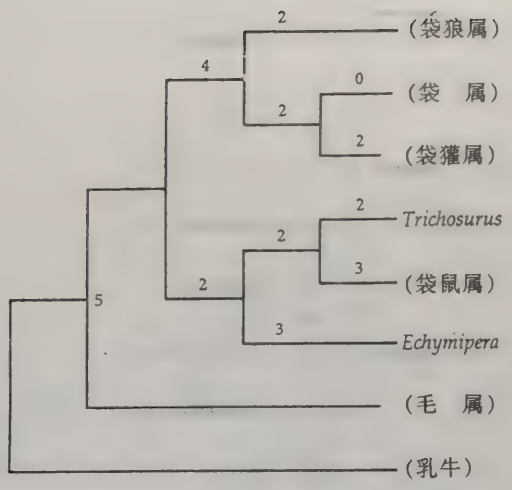


图 5－18　袋狼(*Thylacinus*)和 6 种其他有袋类根据线粒体 12 SrRNA 顺序作出的最节省树。该树用乳牛(*Bos*)作组外单位找到了根。分枝上的数字代表替换数。自 Thomas 等(1989)。

图 5－19　海滩雀(*Ammodramus maritimus*)的 9 个在分类学上有区别的亚种的地理分布。自 Avise 和 Nelson (1989)。

相互间几乎是不可区别的。在此研究中 3 个海湾沿岸的亚种也有同样的关系。相比之下，大西洋沿岸亚种与海湾沿岸亚种间则有相当明显的区别。这两大类群间的每位点核苷酸替换数估计约为 1%。如果海滩雀的线粒体 DNA 以与哺乳类和其他鸟类的线粒体 DNA 大致相同的速率进化(即每百万年 2－4%的顺序分岐)，那么，这两大类群的群体是在大约 25 万－50 万年前发生分离的。虽然这些对分岐的绝对年代的估值，由于缺乏确定的统一标准而停留在定性水平，但它们却与已得到的海平面开始下降的年代一致得相当好。由于海平面下降，佛罗里达半岛露出水面，从而对半岛两侧的群体起了生殖屏障的作用。

更重要的是，阿维斯和纳尔逊(Avise 和 Nelson，1989)的分子研究表明，深色海滩雀亚种与另两个大西洋亚种(即 *A. m. maritima* 和 *A. m. macgillivraii*)在分子上不可区别，而它与海湾亚种，象 *A. m. peninsulae* 却是相当不同的，但繁殖计划中却从后者里选出雌体。结果，深色海滩雀的救援方案也许是建立在错误的系统发育前提上的。该方案可能不是在重建一个灭绝了的亚种，而是创造了一个新亚种。所以，系统发育关系方面的知识在生物多样性的保护中，对作出合理决定是至关重要的。分类学

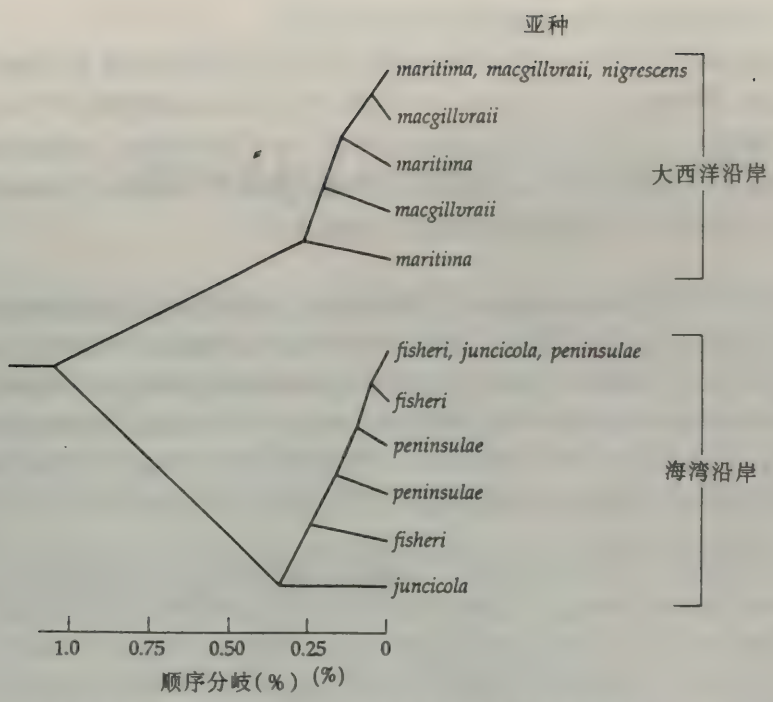


图 5－20　UPGMA 系统树图，表示海滩雀的大西洋沿岸群体和海湾沿岸群体间的线粒体 DNA 基因型区别。应用最节省法可得到许多同样节省的树，其中包括与所示图等同的树。所有供选取的树都涉及大西洋进化枝内部或海湾进化枝内部少量的分枝重排，但这两个类群间的界线却维持不变。系统树图中同样的亚种名多次出现，这表示属于同一亚种的不同个体展现出不同的限制酶模式。反之，几个亚种名在一个分枝的末端处出现则表示，根据形态学和动物地理学而被归于不同亚种的某些个体，对所使用的那些酶表现出相同的限制酶模式。自 Avise 和 Nelson (1989)。

上的失误也许会使那些愿望良好的努力变得劳而无功。

习题

1. 对 4 个 OTU：A，B，C 和 D，画出所有可能的有根和无根树。

2. 找出以下 5 个假想顺序中的所有信息位点：

	1	2	3	4	5	6	7	8
(a)	A	T	G	A	C	T	A	A
(b)	G	T	G	A	T	T	G	A
(c)	A	C	G	G	A	T	A	A
(d)	A	T	G	C	A	T	T	A
(e)	A	C	G	C	A	T	C	A

3. (a)用 UPGMA 法，(b)用变形距离法，和(c)用萨塔特和特韦尔斯基的近邻关系法，对以下距离矩阵构建系统树。在变形距离法的场合，假定 OTU E 对所有其他 OTU 而言是一个已知的组外单位。

	OTU			
OTU	A	B	C	D
B	3			
C	8	7		
D	7	6	3	
E	11	10	13	12

4. 图 5－15b 中的树是用萨塔特和特韦尔斯基的近邻关系法得到的(见第 66-67 页和 71－75

页)。请用第 69-70 页里介绍的方法证明在该树上所标出的枝长。

5. 在经典昆虫分类学中,不完全变态类(如直翅目)和完全变态类(如双翅目、鳞翅目,同翅目)间的分岐曾被假定是非常远古的事件。(a)用图 5－21 中的 6 个 5S rRNA 顺序和 UPGMA 法来构建系统树。(b)用其中 5 个顺序(a、b、c、e 和 f)和最节省法来构建一个无根系统树。把根放在联结组外单位 *Artemia salina*(f)和其他 OTU 的分枝上。这两种树与经典分类学结果一致吗?这两种构建法结果能得到相同的拓扑图吗? 如果不同,那么造成这种差异的原因是什么?

(a) GCCAACGTCCATACCACGTTGAAAGCACCGGTTCTCGTCCGATCACCGAAGTTAAGCAGC
(b) GGCAACGACCATACCACGTTGAATACACCAGTTCTCGTCCGATCACTGAAGTTAAGCAAC
(C) GCCAACGTCCATACCACGTTGAAAACACCGGTTCTCGTCCGATCACCGAAGTCAAGCAAC
(d) GCCAACGTCCATACCACGTTGAAAACACCGGTTCTCGTCCGATCACCGAAGTTAAGCAAC
(e) GCCAACGACCATACCACGCTGAATACATCGGTTCTCGTCCGATCACCGAAATTAAGCAGC
(f) ACCAACGGCCATACCACGTTGAAAGTACCCAGTCTCGTCAGATCCTGGAAGTCACACAAC

(a) GTCGGGCGCGGTTAGTACTTGGATGGGTGACCGCCTGGGAACCCCGCGTGACGTTGGCA
(b) GTCGGGCGTAGTTAGTACTTGGATGGGTGACCGCTTGGGAACACTACGTGCCGTTGGCA
(c) GTCGGGCGTAGTCAGTACTTGGATGGGTGACCGCCTGGGAACACTACGTGATGTTGGCT
(d) GTCGGGCGCGGTCAGTACTTGGATGGGTGACCACCTGGGAACACCGCGTGCCGTTGGCT
(e) ETCGGGCGCGGTTAGTACTTAGATGGGGGACCGCTTGGGAACACCGCGTGTTGTTGGCC
(f) GTCGGGCCCGGTCAGTACTTGGATGGGTGACCGCCTGGGAACACCGGGTGCTGTTGGCA

图 5－21　来自 6 种节肢动物的 5 S rRNA 基因的 DNA 顺序。(a)蝗虫(*Acheta domesticus*，直翅目)(b)蚜虫(*Acyrthosiphon magnoliae*,同翅目)(c)家蚕(*Bombyx mori*,鳞翅目)(d) 蛾(*Philosamia cynthia*,鳞翅目)(e)果蝇(*Drosophila melanogaster*,双翅目)(f)甲壳动物(*Artemia salina*,甲壳纲)。资料来自 Kawata 和 Ishikawa(1982) ,Morton 和 Sprague (1982),Gu 等(1982),Bagshaw 等(1987) ,Cave 等(1987),以及 Samson 和 Wegnez(1988)。

后继阅读文献

Felsenstein. J. 1988. Phylogenies from molecular sequences: Inference and reliability. Annu. Rev. Genet. 22: 521－565

Goodman. M. (ed.) . 1982. *Macromolecular Sequences in Systematic and Evolutionary Biology.* Plenum. New York.

Hillis. D. M. and C. Moritz (eds.). 1990. *Molecular Systematics.* Sinauer Associates. Sunderland. MA.

Margulis. L. 1981. *Symbiosis in Cell Evolution.* Freeman. San Francisco.

Nei. M. 1987. *Molecular Evolutionary Genetics.* Columbia University Press. New York.

Sneath. P. H. A. and R. R. Sokal. 1973. *Numerical Taxonomy.* Freeman. San Francisco.

Woese. C. R. 1987. Bacterial evolution. Microbiol. Rev. 51: 221－227.

6 由基因重复和外显子混匀造成的进化

最先注意到基因重复在进化中的重要性的是霍尔丹(Haldane,1932)和马勒(Muller,1935)。他们认为,一个基因的多余复本也许能发生引起岐化的突变,因而最终将会以一个新基因的形式出现。大野(Ohno,1970)以分子的、生物化学的和细胞学的证据为凭,把这种看法引向了极端,主张基因重复是唯一能引起新基因产生的途径。虽然,现在已经知道还有一些别的产生新功能的方式(见第 95-96 页),但大野的观点在很大程度上还是成立的。

断裂基因的发现启发了吉尔伯特(Gilbert,1978),于是他提出,内含子间的重组为基因间外显子序列交换提供了一种机制。已经发现的许多这类外显子交换的例子表明,这种机制在真核生物的基因以出现新功能的形式进化中,起着十分显著的作用。

6.1 DNA 重复的类型

一个 DNA 片段的拷贝数增加可由几种类型的 DNA 重复(DNA duplication)所引起。这通常根据所涉及的基因组区域的幅度来分类。已经知道有以下几种类型的重复:(1)部分基因重复或基因内重复(partial or internal gene duplication),(2)全基因重复(complete gene duplication),(3)部分染色体重复(partial chromosomal duplication),(4)非整数倍重复或染色体重复(aneuploidy or chromosomal duplication),和(5) 多倍性重复或基因组重复(polyploidy or genome duplication)。前 4 种类型又称区域性重复(regional duplication),因为它们影响的不是整个单倍的染色体组。大野(Ohno,1970) 曾极力主张,基因组重复一般要比区域性重复更为重要一些,因此在后一种情况下,结构基因的调节系统可能只有部分发生了重复,而这种不平衡可能会破坏重复基因的正常功能。然而,正如以下所讨论的,区域性重复显然在进化中也起着非常重要的作用。

DNA 重复长期以来一直被认为是造成基因组大小进化的一个重要因素(见 Ohno,1970)。特别地,全基因组重复或它的某一主要部分,如一条染色体的重复,可能会造成基因组大小突然而极大的增长。基因组重复事件曾在各种不同的生物类群的进化中反复地被记录到,而在植物、真骨鱼类和两栖类中最为突出。造成基因组扩大的进化途径将在第八章中进行讨论。

6.2 域和外显子

一个蛋白质域(domain)是蛋白质中一个定义明确的区域,它区别于蛋白质中的其它部分,或者执行某一特殊功能,如与基质结合,或者构成该蛋白质内的一个稳定、紧密的结构单位,前者称为功能域(functional domain),后者则称结构域(structural domain)或组件(module)(Gō 和 Nosaka,1987)。定义一个功能域的边界常常是很困难的,因为在许多情况下功能是由散布在整个多肽里的氨基酸残基执行的。另一方面,一个结构组件则是由一段连续的氨基酸片段所构成的。

在考虑产生多重域蛋白质的可能进化机制时,以上区别是相当重要的。如果一个功能域相当于一个组件,那么,它的重复将会增加功能片段的数目。反之,如果功能是由散布在不同组件中的氨基酸残基执行的,则一个组件重复所造成的影响也许从功能上看是不成气候的。在许多蛋白质中看到的内部重复常常对应于结构组件或者单组件的功能域(Barker 等,1978)。

从理论上讲,结构域和外显子在基因中的排列间也许能设想出几种可能的关系(图 6-1)。乡(Gō,1981)发现,在许多内部结构域划分已经确定的球状蛋白质中,基因的外显子和该域之间或多或

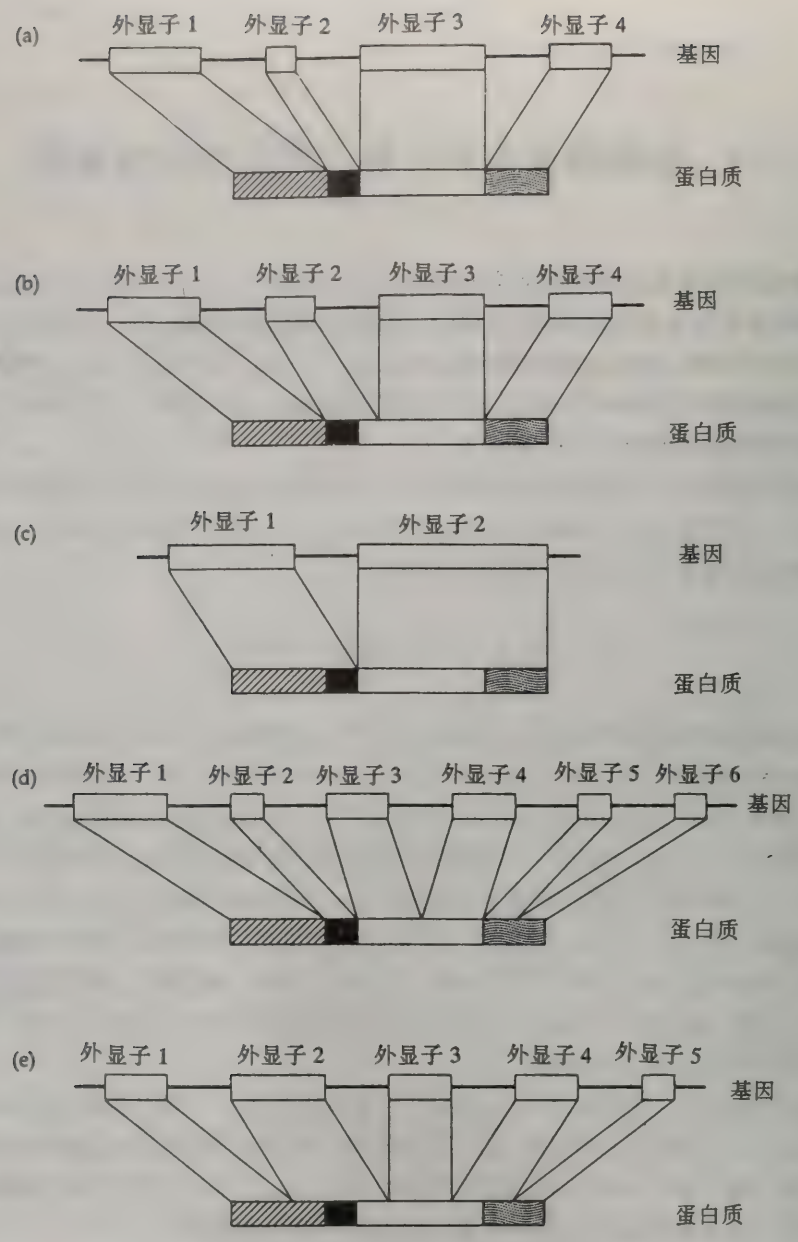


图 6－1　基因中外显子的排列与它所编码的蛋白质的结构域之间的可能关系：(a)每一外显子正好对应于一个结构域；(b)仅近似地对应；(c)一个外显子为两个或更多的域编码；(d)一个结构域由 2 个或多个外显子编码；和(e)外显子和域之间不对应。该蛋白质的 4 个结构域用不同的矩形块(画有斜条纹的、黑色的、白色的和打上点的)表示。

少地存在着精确对应(图 6－1a、b)。在某几个例子中，可看到一个组件由一个以上的外显子编码的现象图(6－1d)。在她的研究中，没有发现一个蛋白质的组件结构与其基因的外显子划分间完全不一致的情况(图 6－1e)。然而，在为数不少的例子中，却可以看到几个邻近的域是由同一个外显子编码的(图 6－1c)。例如，血红蛋白 α 和 β 分别由 4 个域构成，而它们的基因却只分别由 3 个外显子所组成，其中第 2 个外显子则为 2 个邻近的域编码。乡认为，由于两个外显子间的内含子丢失，结果出现了两外显子的合并。事实上，存在于植物中的同源蛋白质——豆血红蛋白，其基因中就可看到在由珠蛋白的域结构预测的位置处(第 68 个氨基酸之后)正好含有一个额外的内含子。所以，珠蛋白基因家族进化期间，有的谱系失去了一些或者全部内含子(图 6－2)。

在大多数情况下，蛋白质水平上的域重复常指示出在 DNA 水平上出现了外显子重复。所以，它表明外显子重复是内部重复的最重要类型之一。真核生物的基因一般由许多外显子和内含子组成(第一章)，而相邻的外显子常常是等同的或相互间非常相似的。这些事实表明，现代生物中许多复合基因

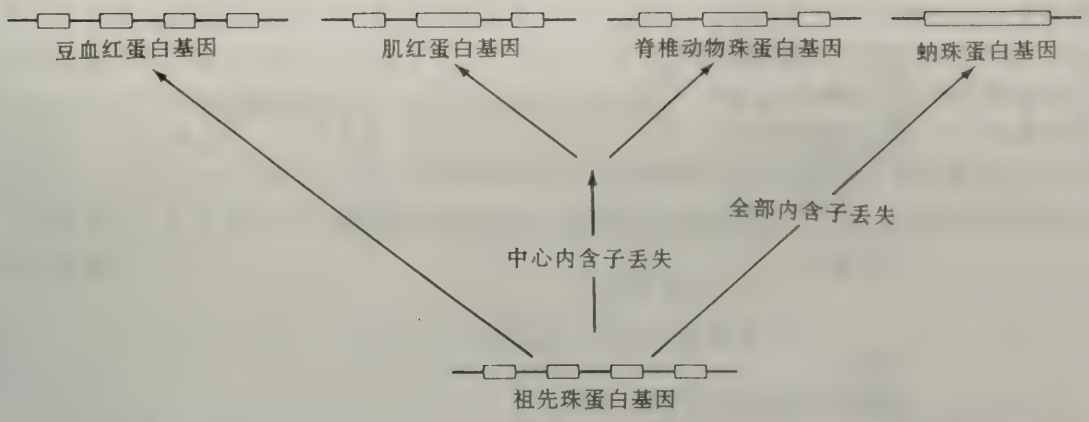


图 6－2　珠蛋白基因进化期间的内含子丢失。原始珠蛋白基因有 3 个内含子和 4 个外显子。豆血红蛋白基因保留着祖先结构，而其他谱系则至少丢失了一个内含子。注意，内含子并未按比例大小画。哺乳动物（牛，人，小鼠，猪和海豹）的肌红蛋白基因中的两个内含子长各为～4800bp 和～3400bp，而珠蛋白和豆血红蛋白基因的同源内含子则分别只有 108－192bp 和 103－904bp 长。豆血红蛋白基因的中间内含子为 99－234bp 长（Blanchetot 等，1983）。珠蛋白的资料来自许多两栖类，鸟类和哺乳类。豆血红蛋白的资料来自 3 种豆类（*Phaseolus vulgaris*，*Glycine max* 和 *Vicia faba*）。

是通过原始基因的内部重复和随后的修饰进化而来的。这类原始基因假定只含 1 个或少数几个外显子，且只能执行简单的生物学功能（Li，1983）。

6.3　域重复和基因的延长

对真核生物的现有基因的勘测表明，内部重复在进化中是频繁发生的。这种在基因大小上的增加，或基因的延长（gene elongation），是简单基因向复合基因进化中最重要的步骤之一。理论上基因的延长也可通过其他方式发生。例如，将一个终止密码子转变成一个有意义的密码子的突变也能使基因延长（第一章）。类似地，一个外来 DNA 片段插入某一外显子中，或出现删除拼接位点的突变，也能得到同样的结果。不过，这类分子变化大多数将破坏延长后的基因的功能，因为加进去的区域是由几乎随机排列的氨基酸所构成的。事实上，在绝大多数情况下，这类分子变化是与病理学表象一起而被发现的。例如，异常血红蛋白恒春（Constant Spring）和 Icaria 分别是由将终止密码子变成谷氨酰胺和赖氨酸的突变所引起的。由于这种突变，这些变异型的 α 链上增加了 30 个残基（Weatherall 和 Clegg，1979）。相比之下，一个结构域的重复倾向不会造成这类问题。事实上，这类重复有时甚至能加强新产生的蛋白质的功能，例如，增加活性位点的数目即可达到这一点。

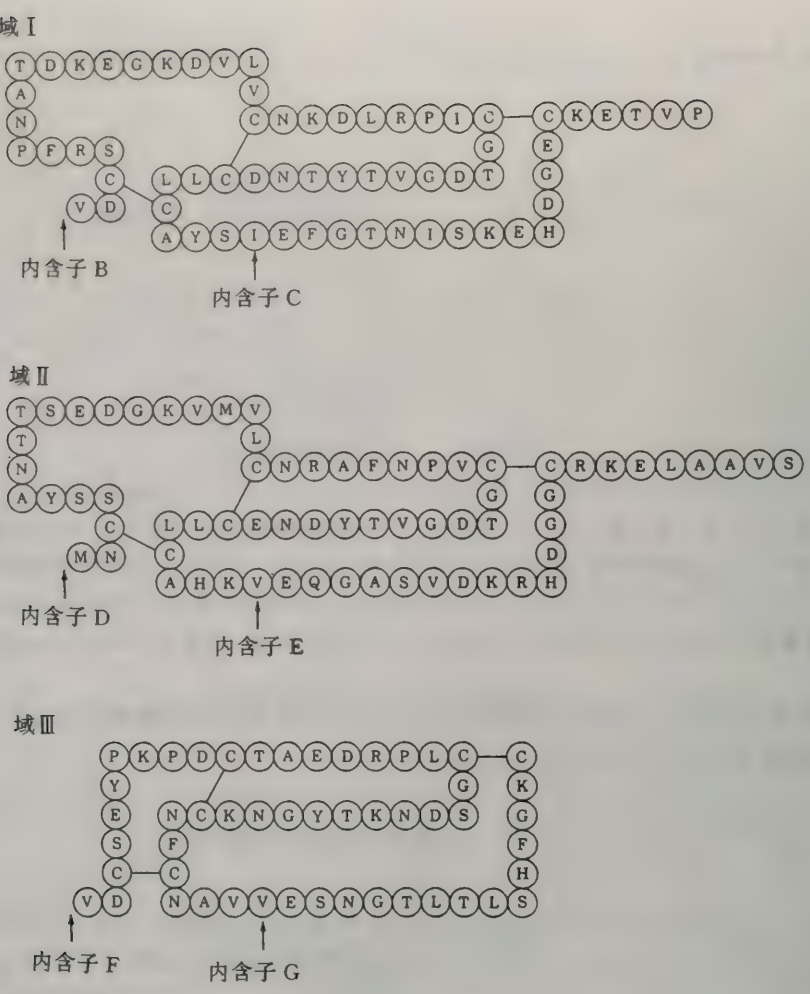
所以，进化期间基因的延长看来主要靠域的重复来实现。在下节中，我们将给出一个基因内重复的例子，以说明进化期间基因延长的后果。

卵类粘蛋白基因

卵类粘蛋白是一种存在于鸟类的卵白中的蛋白质，它能抑制一种催化蛋白质分解的胰蛋白酶的活性。卵类粘蛋白多肽可被划分成 3 个功能域（图 6－3）。每一个域都能和一个分子的胰蛋白酶或其他丝氨酸类蛋白酶结合。为这 3 个功能域编码的 DNA 区域明显地有着共同的进化起源，且相互间由内含子所隔开（Stein 等，1980）。这 3 个区域中，每一个都是由被一个内含子隔断的两个外显子所构成，且这两个外显子间不表现出相似。于是，卵类粘蛋白基因看来是由一个原始的单域基因经两次内部重复而得来的，其中每次重复都涉及两个邻近的外显子。由于域Ⅰ和Ⅱ相互间比它们中的任一个和域Ⅲ之间都更为相似，所以它们可能是经第 2 次重复而得到的，而域Ⅲ则是第 1 次重复的产物。

域重复的普遍性

表 6－1 列出了几个证据表明它们在其进史中发生过内部重复的基因名单。这些基因全



域	相似性百分数	
	氨基酸	核苷酸
Ⅰ vs. Ⅱ	46	66
Ⅱ vs. Ⅲ	30	42
Ⅰ vs. Ⅲ	33	50

图 6－3　分泌性卵类粘蛋白的3个功能域，和域间氨基酸水平与核苷酸水平的顺序相似程度。自 Stein 等，(1980)。

表 6－1　具有内部域重复的蛋白质

序　列	蛋白质的长度[a]	重复的长度	重复的次数	重复的百分比[b]
免疫球蛋白∈-链 C 区(人)	423	108	4	100
免疫球蛋白 γ-链 C 区(人)	329	108	3	98
血清白蛋白(人)	584	195	3	100
小白蛋白(人)	108	39	2	72
蛋白酶抑制因子，Bowman-Birk 型(大豆)	71	28	2	79
蛋白酶抑制因子，颌下腺型(啮齿类)	115	54	2	94
铁氧还蛋白(*Clostridium pasteurianum*)	55	28	2	100
血纤蛋白溶酶原(人)	790	79	5	50
钙依赖性调节蛋白(人)	148	74	2	100
原肌球蛋白 α 链(人)	284	42	7	100

自 Barker 等(1978)

a. 氨基酸残基数。b. 由重复顺序占据的部分占蛋白质总长的百分比。

都涉及一个或多个域重复，而其中有些序列则是由一个原始序列经多次重复而得到的，结果使这种重复性结构占据了整个蛋白质的长度。在这些例子中，每一例的重复事件都可从蛋白质或 DNA 顺序的类似而轻易地推测出来。也许还有许多别的复合基因也是经基因内重复而进化的，但它们的重复区域相互间可能已分岐到这样一种程度，以致它们间的顺序同源性已经不能被辨认出来了。在某些情况下，如免疫球蛋白基因的恒定区和可变区，我们可以通过比较这些域的二级结构来推测其共同祖先，因为二级结构有比氨基酸顺序更强的保守性。所以，蛋白质中的内部重复极有可能比经验数据所指示的更多地普遍地存在着。

6.4 基因家族的形成与新功能的获得

一次全基因重复产生两个等同的拷贝。它们将如何进化则会因情况而异。例如，这些拷贝可能保留其原始功能，从而使该生物产生更多的某种 RNA 或蛋白质。此外，其中一个拷贝可能会因一次有害突变而丧失能力，从而变成一个无功能的假基因（见第 89 页）。然而，更重要的是，基因重复可能会导致产生遗传新型或新基因的结果。如果重复中一个拷贝保留其原始功能，而另一个则累积分子变化，以致于最终变得能执行完全不同的功能，那么，产生遗传新型或新基因的情况就会出现了。

重复的基因可以分成两类：变异的重复和不变的重复。不变的重复（invariant repeats）相互间在顺序上是等同的或近似地等同的。在有些情况下表明，等同顺序的重复与某一基因产物的增量合成有关，该产物则是生物的正常功能所必需的。这样的重复称为剂量重复（dose repetitions）。无论何时出现需产生大量特别的 RNA 或蛋白质产物的代谢需要，剂量重复就会十分普遍的出现（Ohno，1970）。代表性的例子有：执行翻译功能不可缺少的 rRNA 的基因和 tRNA 的基因，以及染色体首要结构蛋白，组蛋白的基因，因此必须被大量地合成。

变异的重复（variant repeats）由一个基因的多拷贝所构成，虽然这些拷贝相互类似，但在其顺序方面却或多或少地有一定程度的差异。有趣的是，变异的重复有时能执行显然不同的功能。例如，在血液凝结过程中起裂解血纤蛋白原作用的凝血酶，和消化性酶胰蛋白酶，都是源自一个原始基因的重复。类似地，乳清蛋白，催化乳糖合成的酶的一个亚基，和通过裂解某些细菌细胞壁中的多糖成份来溶解它们的溶菌酶，在谱系上是相关的。功能上的分化通常需要大量的替换。不过，在某些情况下，一个新的功能有可能在数目相对较小的替换之后产生（例如，见 Betz 等，1974）。

在一个基因组中属于某一群重复顺序的所有基因，合起来被称为一个基因家族（gene family）或多基因家族（multigene family）。一个基因家族的成员通常位于同一染色体上相互间极靠近的地方。在某些情况下，一些功能性的或非功能性的家族成员可能会位于别的染色体上。

当重复基因在功能或顺序上变得相互间差异很大时，再把它们归成同一基因家族也许就不合适了。戴霍夫（Dayhoff，1978）造了一个词：超家族（superfamily）来描绘关系密切和关系疏远的蛋白质间的联系。据此，在氨基酸水平上相互至少展示出 50％相似性的蛋白质，可以被看成是一个家族的成员；而当同源蛋白质展示出的相似性小于 50％时则被看成是一个超家族的成员。例如，α－珠蛋白和 β－珠蛋白被分类在两个不同的家族中，而它们和肌红蛋白一起则构成了珠蛋白超家族（见第 99 页）。然而，这两术语并不总是能按戴霍夫的标准来严格地应用的。例如，人和鲤的 α 珠蛋白链仅展现出 46％的顺序相似性，这就低于为归于同一基因家族所划的界限。为此缘故，将蛋白质归类成家族和超家族，不仅要根据顺序的相似性，而且还要考虑关于功能类似性或组织特异性等方面的辅助证据才能决定。

基因家族内的基因数变化极大。有些基因仅在基因组内重复几次，它们被称为轻度重复（lowly repetitive）。另一些则可能在基因组中上百次地重复，因而被称为是高度重复的（highly repetitive）。在以下几节中，rRNA 和 tRNA 基因将被作为例子，以说明高度重复的不变基因。轻度重复的基因将由同功酶和色敏感的色素蛋白基因作代表。

确定 RNA 的基因

表 6－2 列出了几种有机体的 rRNA 和 tRNA 基因的数目。哺乳动物的线粒体基因组只含有一个

拷贝的 12SrRNA 基因和一个拷贝的 16SrRNA 基因。这对线粒体翻译系统来说显然是足够了，因为其基因组仅含 13 个为蛋白质编码的基因(第四章)。枝原体是最小的自我复制的原核生物，它含有两组 rRNA 基因。大肠杆菌的基因组大小是它的 4—5 倍，含有 7 组 rRNA 基因。酵母中 rRNA 基因的数目大约是 140，果蝇中和人中的数目则更大。爪蟾 *Xenopus laevis* 有比人更大的基因组和更多的 rRNA 基因。所以，rRNA 基因的数目与基因组大小间存在着很强的正相关。这一规则对 tRNA 基因(表 6—2)和其他确定 RNA 的基因来说也是成立的。

表 6—2　各种有机体中每单倍体基因组的 rRNA 和 tRNA 基因数

基因组来源	rRNA 基因的数目[a]	tRNA 基因的数目	基因组的近似大小(bp)
人线粒体	1	22	1.7×10^4
枝原体 *Mycoplasma capricolum*	2	ND[b]	1×10^6
大肠杆菌 *E. coli*	7	～100	4×10^6
酵母 *Saccharomyces cerevisiae*	～140	320—400	5×10^7
果蝇 *D. melanogaster*	130—250	～750	2×10^8
人	～300	～1300	3×10^9
爪蟾 *Xenopus laevis*	400—600	～7800	8×10^9

自 Li(1983)

a. 对于 rRNA 基因，该值指整个 rRNA 基因组的数。

b. ND＝未定。

高度重复的基因，象 rRNA 基因，一般相互间是非常相似的。造成这种同源性的一个因素可能是纯洁化选择，因为这些基因应该遵守非常特殊的功能和结构要求。然而，同源性常常会延伸到没有任何功能意义的区域，而这类同源性的维持就需要求助于别的机制(见第 99 页)了。

同功酶

除了不变的重复以外，高等生物的基因组还含有许多其成员相互间已发生不同程度岐化的多基因家族。其中最能说明问题的例子是为同功酶编码的基因家族，象乳酸脱氢酶、醛缩酶、肌酸激酶和丙酮酸激酶等。同功酶(isozymes) 是催化同样的生化反应、但在组织特异性、生长调控、电泳移动性或生化特性等方面相互间可能有差别的一类酶。注意，同功酶是由不同的基因座位，通常是重复后的基因来编码的，这与异型酶(allozymes)不同，后者是由同一个基因座位上的不同等位基因编码的、某种酶的不同形式。

让我们来考虑为脊椎动物中乳酸脱氢酶(LDH)的 A 和 B 亚基编码的两个基因。这两种亚基可形成 5 种四聚体同功酶：A_4，A_3B，A_2B_2，AB_3 和 B_4，所有这些酶都可在氧化态辅酶，尼克酰胺腺嘌呤二核苷酸(NAD^+)的存在下催化将乳酸转变成丙酮酸的反应，或在还原态辅酶(NADH)的存在下催化相反方向的反应。曾经有人提出，LDH—B_4 和另外一些富含 B 亚基的同功酶对 NAD^+ 有较高的亲和性，它们在一些行有氧代谢的组织如心脏中，行使着真正的使乳酸脱氢酶的功能，而 LDH—A_4 和另一些富含 A 亚基的同功酶则对 NADH 有较高的亲和性，所以，它们在一些无氧代谢的组织如骨骼肌里，被特别地安排作为丙酮酸还原酶而起作用(Everse 和 Kaplan，1975；Nadal-Ginard 和 Markert，1975)。图 6—4 展示了心脏中产生的 LDH 的发育次序。我们看到，心脏存在的环境越是厌氧，特别是在怀孕的早期阶段，则富含 A 亚基的 LDH 同功酶所占的比例就越高。于是，这两个重复基因已对不同的组织和不同的发育阶段发生了特化。因为这两种亚基存在于几乎所有曾做过年代测定研究的脊椎动物中，所以，产生 LDH—A 和 LDH—B 的基因重复可能发生在脊椎动物进化的早期阶段之前或期间。LDH 的一个有趣特 性是，这两个亚基能形成异源多聚体，从而进一步增加了该酶的生理学功能多样性。

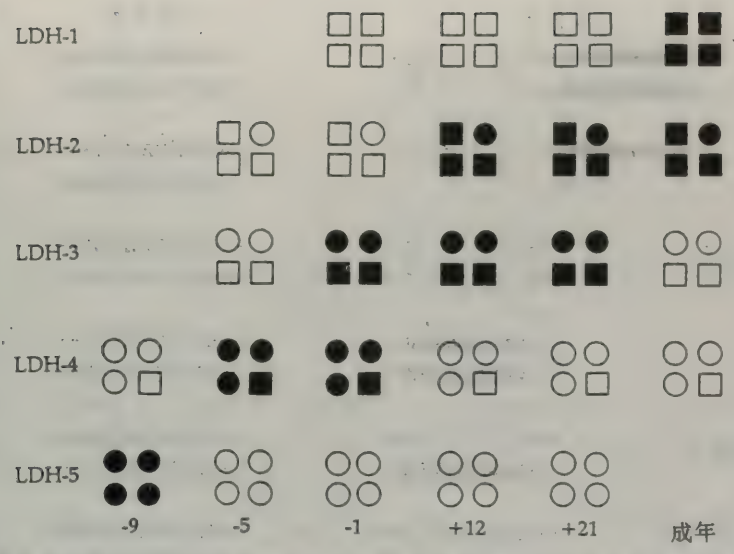


图 6－4　心脏中 5 种乳酸脱氢酶(LDH)同功酶的发育次序。负数和正数分别表示出生前、后的天数。方块表示 B 亚基、圆圈表示 A 亚基。被涂黑了的图形表示在数量上占优势的形式。注意在个体发育期间从 A 亚基向 B 亚基的转移。资料取自 Markert 和 Vrsprung(1971)。

色敏感色素蛋白

人、猿和古世界猴具有 3 种色敏感色素蛋白。蓝色素由一个常染色体基因编码，而红色素和绿色素则各由一个 X-连锁基因编码(Nathans 等，1986)。红和绿色素的氨基酸顺序有 96%是等同的，但它们与蓝色素的相似性却只有 43%。蓝色素基因和绿、红色素基因的祖先在大约 5 亿年前发生分歧。相比之下，红和绿色素间的紧密连锁和高度的同源性指出，它们来源于非常近期的基因重复。因为新世界猴只有一个 X-连锁的色素基因，而古世界猴和人则有 2 个或多个色素基因，所以可以假定，重复发生在大约 3500－4000 万年以前、古世界猴与新世界猴分岐以后的祖先中。作为该重复的结果，人、猿和古世界猴能分辨 3 种颜色(即它们是三色性的)，而新世界猴，如松鼠猴，则只能对蓝与绿或蓝与红加以区别，但不能对绿与红加以区别(即它们是二色性的)。

有趣的是，两个 X-连锁的等位基因呈杂合状态的雌性松鼠猴是三色性的(Jacobs 和 Neitz，1986)。另一方面，只携带一条 X 染色体的雄体则从未获得过三色性视觉。于是，在人和古世界猴的场合，三色性视觉是通过类似于同功酶的机制(即，两个有区别的蛋白质由不同的基因座位编码)而获得的，而杂合体雌性松鼠猴达到同样目的，则是通过应用两种异型酶(即，同一基因座位上两个不同的等位基因形式)来实现的(图 6－5)。如果三色性视觉能 给予其携带者以选择优势，那么，新世界猴的一基因座位上的两个色敏感等位基因的长期维持，可能是通过一种超显性选择形式来实现的(第二章)。

6.5　重复基因的无功能化

多余的重复基因更可能是变成无功能基因、而不是进化成一个新基因，因为有害突变远比有利突变发生得频繁。一个重复基因的无功能化即产生一个假基因。这样产生的假基因称未加工的(unprocessed)假基因，这与将在第七章中讨论的经过加工的假基因相反。表 6－3 列出了在几种珠蛋白假基因中发现的结构缺陷。这些未加工的假基因大多数含有多重缺陷，象阅读框架移动、成熟终止前终止、和拼接位点或调控位点的删除等，以至很难看出哪种突变是使基因沉默的直接原因。在有几个例子里，也许能找到“元凶”。例如，人的 $\psi\zeta$ 只含有一个严重缺陷，无义突变，所以它可能是无功能化的直接原因。(符号 ψ 用来将假基因与其有功能的对应物加以区别。)有些假基因，象在山羊 β－珠蛋白多基因家族中的 $\psi\beta^x$ 和 $\psi\beta^z$，则是由一个预先存在的假基因重复而得来的。

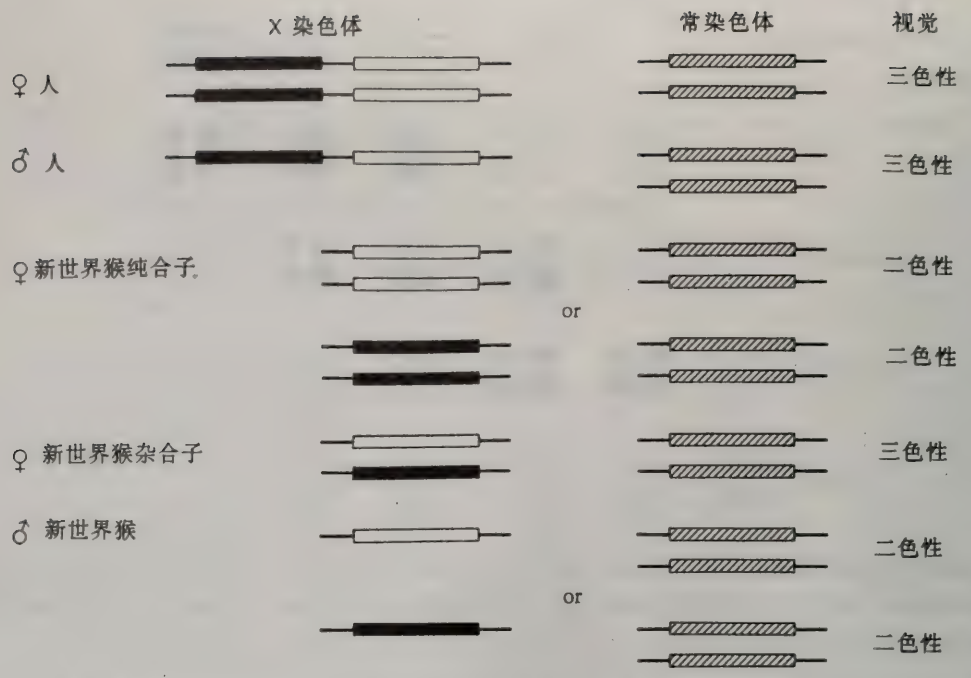


图 6—5　人和新世界猴(NWM)的雄性和雌性中，三色性视觉的分子基础。注意，雄性新世界猴不能获得三色性视觉。涂黑了的、空心的和画斜线的矩形分别表示绿、红和蓝色素基因。

表 6—3　珠蛋白假基因中的缺陷[a]

假基因	TATA 框	起始密码子	阅读框架移动	成熟终止前终止	缺乏必需氨基酸	拼接 GT/AG 规则	改变了的终止密码子	多聚腺苷化信号 AATAAA
人 ψα1		+	+	+	+	+	+	+
人 ψζ1				+				
小鼠 ψα3	+		+	+		+		
小鼠 ψα4			+		+			
小鼠 βh3	?	+	+	+	+	+	?	?
山羊 ψβ[x]	+		+	+	+	+	+	+
山羊 ψβ[y]	+		+	+	+	+	+	+
兔 ψβ2			+	+	+	+		

自 Li(1983)

a、加号表示存在一种特别类型的缺陷；问号表示有存在该缺陷的可能性。

6.6　基因重复的年代测定

两个基因，如果它们是从一次重复事件中得来的则称为平行相关的(paralogous)，如果它们是从一次物种形成事件中得来的则称为垂直相关的(orthologous)。例如在图 6—6 中，基因 α 和 β 是从一个祖先基因的重复中得到的，因而是平行相关的，而来自物种 1 的基因 α 和来自物种 2 的基因 α 则是垂直相关的，来自物种 1 的 β 基因和来自物种 2 的 β 基因间的关系也是如此。

如果我们知道基因 α 和基因 β 中的替换速率，则我们即可从序列资料中估出重复的年代，即 T_D。而替换速率则可根据垂直相关的基因间的替换数，结合有关物种 1 和物种 2 间分岐时间 T_S 的知识(图 6—6)来估出。下面我们将说明 T_D 的估值是怎样得到的。

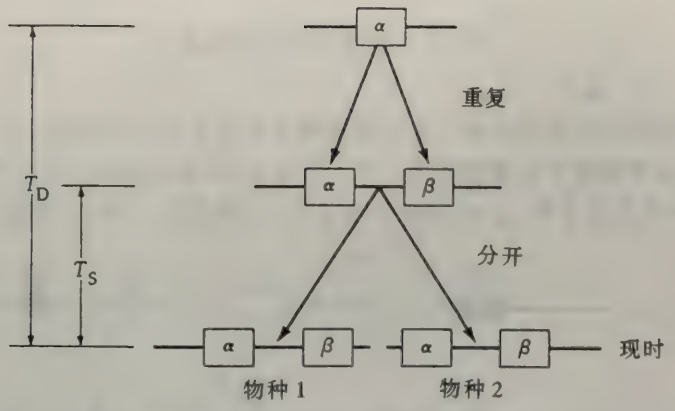


图 6—6　用来估计一个基因重复事件的时间(T_D)的模型。α 和 β 这两个基因来自 T_D 单位时间以前一个祖先物种中发生的重复事件。后来该物种分裂成两个物种,1 和 2,这发生在 T_S 时间单位以前。在物种 1 和物种 2 中的两个 α 基因是垂直相关的,两个 β 基因也是如此,但 α 基因和 β 基因间则是平行相关的。

对于基因 α,设 K_α 为这两个物种间的每位点替换数。那么,在基因 α 中的替换速率则由

$$r_\alpha = \frac{K_\alpha}{2T_s} \tag{6.1}$$

来估计。在基因 β 中的替换速率,r_β 可用同样方式得到。这两个基因的平均替换速率则为:

$$r = \frac{r_\alpha + r_\beta}{2} \tag{6.2}$$

为了估出 T_D,我们需要知道基因 α 和 β 间的每位点替换数($K_{\alpha\beta}$)。这个数可从以下 4 个对子的比较中得到:(1)来自物种 1 的基因 α 和来自物种 2 的基因 β,(2)来自物种 2 的基因 α 和来自物种 1 的基因 β,(3)来自物种 1 的两个基因,(4)来自物种 2 的两个基因。从这 4 个估值我们能算出 $K_{\alpha\beta}$的平均值($\overline{K_{\alpha\beta}}$),进而我们能估出 T_D 为:

$$T_D = \frac{\overline{K_{\alpha\beta}}}{2r} \tag{6.3}$$

注意,在蛋白质编码基因的情况下,分别用同义替换数和非同义替换数,我们能得到两个相互独立的 T_D 估值。这两个估值的平均值也许能作为 T_D 的最后估值来使用。然而,如果基因 α 和 β 间每同义位点的替换数太大,比如说大于 1,则同义替换数就不能被精确地估出,这样同义替换也许不能提供一个可靠的 T_D 估值。在这种情况下,将只有非同义替换数得到应用。反之,如果这种平行相关的基因间每非同义位点的替换数太小,那么,非同义替换数的估值将承受较大的取样误差,在这种情况下,就只应该用同义替换数了。

以上我们是在速率恒定的假定下进行的。此假定可用以上提及的 4 个对子的比较来加以检验。如果这 4 个对子的比较中近似相等性不成立,则该假定也不能成立。如后面(见第 95 页)将要讨论的那样,起因于具体的进化事件的一些问题也可能会产生,并使 T_D 的估计复杂化。

测定基因重复事件年代的另一种方法是,结合有关被研究物种分岐年代的古生物学资料,来考虑基因在系统发育中的分布。例如,除无颌鱼(Agnatha)外所有脊椎动物都编码 α 和 β 珠蛋白链。对此观察到的事件有两种可能的解释。一种是:产生 α 和 β 珠蛋白的重复事件发生在无颌类与其他脊椎动物的共同祖先中,但后来所有无颌类都丢失了其中的一个重复。这是有可能的但可能性不大,因为这样一种设想需要在许多进化谱系中发生的这种丢失是非独立的(即与物种有关)。另一种解释是:重复事件发生在无颌类与其他所有脊椎动物的祖先分岐之后,但在其他脊椎动物幅射演化之前(4.5—5 亿年前)。后一种解释想来要更合理一些,所以重复的年代普遍取 4.5—5 亿年前(Dayhoff,1972;Dickerson 和 Geis,1983)。

显然,以上方法只能粗略给我们提供重复年代的估值,所以,对所有估值都应该小心采用。

6.7 珠蛋白基因超家族

珠蛋白超家族曾经历过所有可能发生在重复顺序家族中的进化路线：(1)原始功能的保留，(2)新功能的获得，和(3)某些重复中功能的丧失。对人类而言，珠蛋白超家族由3个家族构成：肌红蛋白家族，它的唯一的一个成员位于第22染色体上，位于第16染色体上的α—珠蛋白家族，和位于第11染色体上的β—珠蛋白家族(图6—7)。这3个家 族合在一起产生两种功能蛋白质：肌红蛋白和血红蛋白。这两种蛋白质在约6—8亿年前发生分岐(见图6—8；Dayhoff，1972；Doolittle，1987)，并已经在某些方面发生了特化。从组织特异性角度看，肌红蛋白变成了肌肉中的储氧蛋白质，而血红蛋白则变成了血液中氧的运输员。就四级结构而言，肌红蛋白保留着单体性结构，而血红蛋白则变成了四聚体。从功能上看，肌红蛋白已进化为有比血红蛋白更高的氧亲合力，而血红蛋白的功能则变得更加精密而可调节(见 Stryer，1988)。例如，哺乳类的血红蛋白，有根据血液中有机磷酸盐的水平来调节其对氧的亲合性的能力。显然，异源多聚体结构有助于血红蛋白功能的精密化。

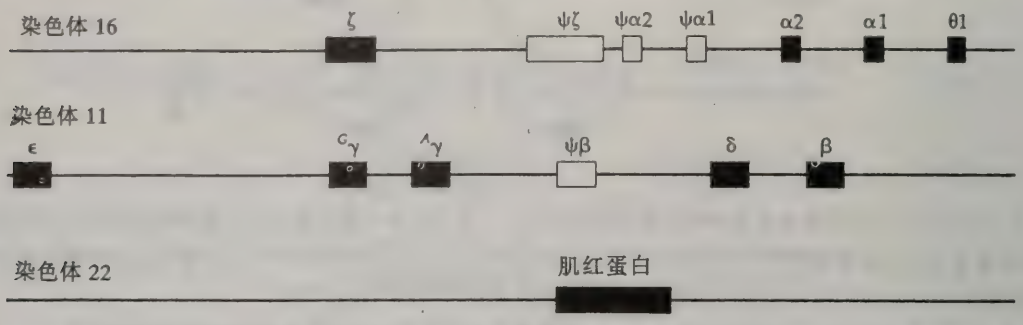


图6—7 人的珠蛋白基因超家族的3个基因家族的染色体排列：α 珠蛋白家族在第16染色体上，β—珠蛋白家族在第11染色体上，而肌红蛋白则在第22染色体上。涂黑了的矩形块表示有功能的基因，空心矩形块则表示假基因。

人和绝大多数脊椎动物的血红蛋白是由两种链所构成，一种由α家族的成员编码，另一种则由β家族的成员编码。如以上所讨论的，α家族和β家族是在约4.5—5亿年前发生分岐的(图6—8)。由于无颌鱼只具有一种单体的血红蛋白，所以，脊椎动物血红蛋白的多聚体化一定是在靠近α—β分岐的时刻出现的。

在人类中，α家族由4个功能基因，即ζ，α_1、α_2和最近发现的θ1所构成(图6—7)。它还含有3个假基因：ψζ，ψα1和ψα2。β家族由5个功能基因：即ε，$^{G}\gamma$，$^{A}\gamma$，β和δ，以及一个假基因ψβ所组成。这两上家族已经在生理学特性和个体发育调节等两个方面都发生了分岐。事实上，在不同的发育阶段上有不同的珠蛋白出现：胚胎期为$\zeta_2\varepsilon_2$和$\alpha_2\varepsilon_2$，胎儿期为$\alpha_2\gamma_2$，而在成年期则为$\alpha_2\beta_2$和$\alpha_2\delta_2$；θ1在何时表达尚且不知。而且，与氧结合的亲合性方面的差别，也在这些珠蛋白中发生了进化。例如，胎儿血红蛋白$\alpha_2\gamma_2$有比任何一种成人血红蛋白($\alpha_2\beta_2$和$\alpha_2\delta_2$)都高的氧亲合力，因而能在处于相对缺氧(低氧)环境的胎儿中更好地行使功能(Wood 等，1977)。这一现象再次为这样一个事实作出了例证，即基因重复能导至生理系统精细化的结果。

在α家族的成员中，胚胎型ζ是分岐程度最高的，在3亿多年以前就已经分枝了(图6—8)。θ1珠蛋白在大约2.6亿年前发生分枝。由于两个α基因间的分岐时间还未确定，所以该图中只画出了α_1基因。α_1基因和α_2基因有几乎等同的DNA顺序，且产生同样的多肽。看起来这好象表示它们的分岐时间离现在非常近。然而，这种类似性也可能是协同进化的结果(Zimmer 等，1980)，这一现象我们放在后面讨论(见第101-102页)。这两个基因存在于人类和所有猿类中，所以可能是在2000多万年以前产生的。

在β家族的成员中，成年型(β和δ)与非成年型(γ和ε)大约在1.55—2亿年前分岐(Efstratiadis 等，1980)。两个γ基因的祖先大约在1—1.4亿年以前与ε基因分岐。产生$^{G}\gamma$和$^{A}\gamma$的重复在大约3500万年前，人谱系与新世界猴谱系分开之后出现(Shen 等，1981)。δ基因和β基因间的分岐以前估计是

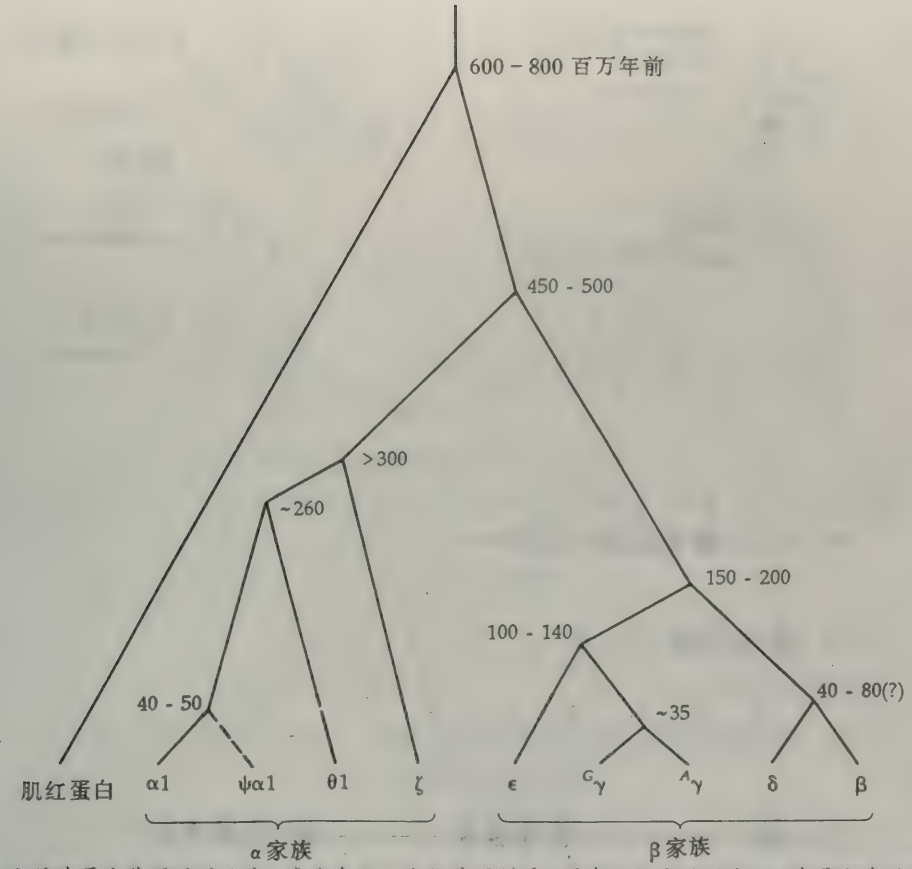


图 6－8　人的珠蛋白基因的进化史。虚线表示一个假基因谱系。图中只标出了两个α－珠蛋白基因中的一个，因为它们相互间发生分歧的年代还未确定。

在 4000 万年前(Dayhoff，1972；Efstratiadis 等，1980)，但最近 DNA 顺序资料表明，它可能早于真兽类的辐射，即在约 8000 万年前出现(Hardison 和 Margot，1984；Goodman 等，1984)。从以上讨论中我们注意到，在两个家族中，基因间分岐时间与基因间功能或调节方面的分岐程度之间，存在着明显的相关。

6.8　外显子混匀

有两类外显子混匀(exon shuffling)：外显子重复和外显子插入。外显子重复指一个基因中的一个或多个外显子的重复，所以它是一种内部重复，这已在基因的延伸一节中讨论过(见第 81 页)。外显子插入是这样一种过程，通过该过程结构域或功能域在蛋白质之间发生交换，或者被插入一个蛋白质之中。这两类混匀都曾在产生新基因的进化过程中被采用。这里，我们将讨论一个外显子从一个基 因插入另一个基因，结果产生镶嵌或嵌合蛋白质的情况(Doolittle，1985；Patthy，1985)。

镶嵌蛋白质

第一个被发现的镶嵌蛋白质是组织血纤蛋白溶酶原活化因子(TPA)(图 6－9)。血纤蛋白溶酶原经 TPA 作用转化成它的活化形式：血纤蛋白溶酶，后者则将血纤蛋白、血块中的一种可溶性的纤维状蛋白质溶解。在底物血纤蛋白的存在下，血纤蛋白溶酶原转化成血纤蛋白溶酶的过程将被大大加速。血纤蛋白能与血纤蛋白溶酶原和 TPA 两者结合，从而将它们联系起来而起催化作用。这种分子排列方式允许血纤蛋白溶酶仅以非常接近血纤蛋白的形式产生，从而给予血纤蛋白溶酶原以对血纤蛋白的特异性。相比之下，尿激酶(UK)，一种尿液中的血纤蛋白溶酶原活化因子，则缺乏血纤蛋白特异性。对 TPA 和 UK 的前体尿激酶原的氨基酸顺序比较表明，TPA 在其氨基末端含有 43 个残基序列，而 UK 中却没有相应的对应物(Banyai 等，1983)。这一片段能形成一种手指样结构(图 6－9a)，而它和另一种蛋白质的与血纤蛋白亲合性有关的指状域是同源的。后一种蛋白质是一种存在于血浆中

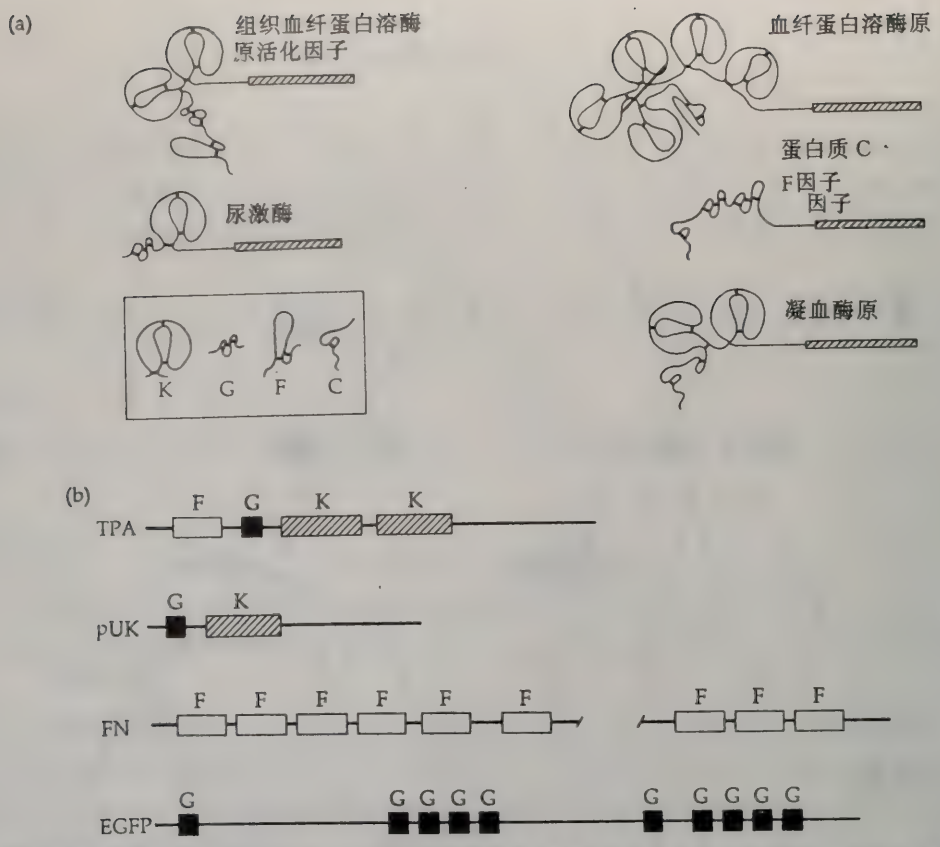


图 6－9 (a)组织血纤蛋白溶酶原活化因子(TPA)与其他涉及血液凝结和血纤蛋白溶解的蛋白质中观察到的各种结构组件。方框插图显示了非蛋白酶区域中组件的结构:K,纽结杆状组件;G,生长因子组件;F,指状组件;C,依赖维生素K的钙结合组件。画有斜线的杆状图形表示与胰蛋白酶同源的蛋白酶区域。自Patthy(1985)修改而成。(b)通过外显子插入组织血纤蛋白溶酶原活化因子(TPA)蛋白质而获得的组件的起源。pUK,尿激酶原,EGFP,表皮生长因子前体;FN,纤粘连蛋白(fibronectin)。

或细胞表面上的、起促进细胞粘着作用的大分子糖蛋白称纤粘连蛋白(FN)(图 6－9b)。若指状域片段缺失会导至 TPA 的血纤蛋白亲合性丧失。TPA 与 FN 的同源性限制在这种指状域内。所以,外显子混匀必定是与 TPA 从 FN 或某一类似蛋白质那里获得该域的事件有关的。

TPA 还含有一个与表皮生长因子(EGF)和另一些蛋白质的类生长因子区域同源的片段。另一些蛋白质是指象因子Ⅸ和因子Ⅹ之类的蛋白质,它们是在血液凝结过程中促使血液结块的酶。此外,TPA 的羧基端区域与胰蛋白酶和其他胰蛋白酶样的丝氨酸蛋白酶的酶蛋白部分同源,这些酶都起着将蛋白质水解成多肽片段的作用。最后,TPA 的非酶蛋白部分含有两个类似于血纤蛋白溶酶原的纽结杆(Kringle)的结构。(一个纽结杆是一个富含半胱氨酸的序列,序列中含有 3 个内部二硫桥并形成一种有点象丹麦蛋糕的纽结杆状结构,这种丹麦蛋糕的名字是(Kringle)。所以,在 TPA 的进化期间,它至少从其他 3 个基因:血纤蛋白溶酶原,表皮生长因子和纤粘连蛋白那里,夺取了至少 4 个 DNA 片段(图 6－9b)。而且,这些得来的单位的连结处和外显子与内含子间的边界精确地一致,于是,就使得关于外显子事实上曾从一个基因转移到另一个基因的想法更加可信。有关外显子混匀的更多例子,可见 Doolittle(1985)和 Patthy(1985)。

外显子混匀上的相位限制

为了使一个插入某一基因的内含子中的外显子不至于引起阅读框架上的框架移动,接受基因时的相位限制 就必须遵守。为了把这种限制搞清楚,让我们根据内含子相对于编码区的可能位置来将内含子归于不同的类型。位于两编码区之间的内含子,根据编码区被打断的方式可被分成 3 种类型。如果内含子处在两个密码子之间,则它是具有相位 0 的内含子;如果它位于一个密码子的第 1 和第 2 个核苷酸之间,则它具有相位 1;如果它位于一个密码子的第 2 和第 3 个核苷酸之间,则它具有相位 2

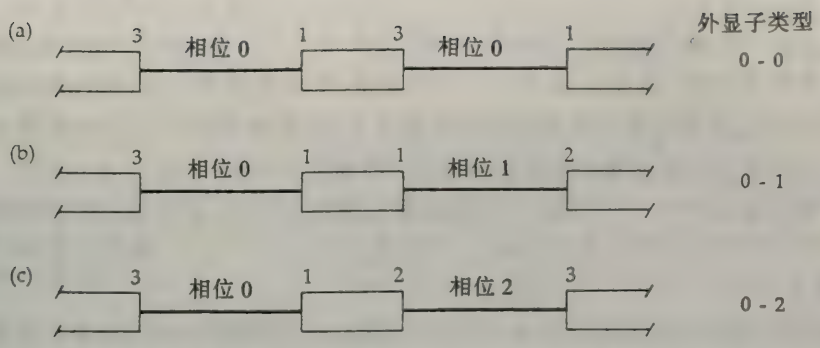


图 6—10　内含子的相位和外显子的类型。外显子用矩形块表示。外显子—内含子连结处上的数字指示外显子的最后一个核苷酸的密码子位置，内含子—外显子连结处上的数字则指示外显子的第一个核苷酸的密码子位置。9 种可能的外显子类型中只有 3 种在图中表示出来了。

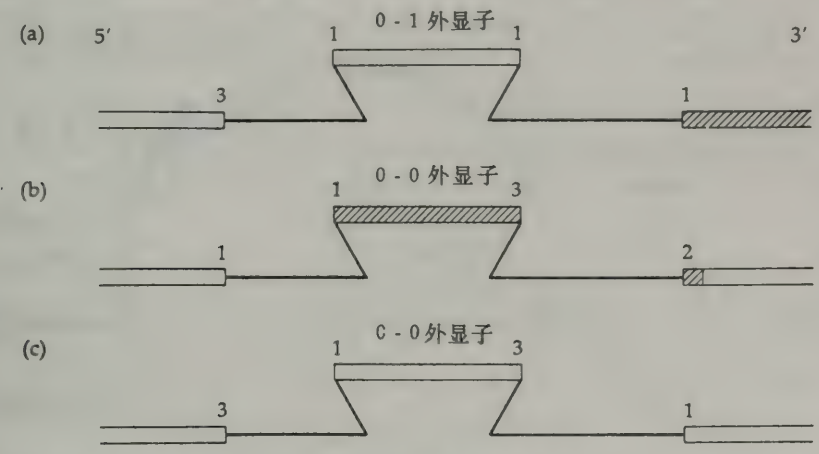


图 6—11　外显子插入内含子的后果。画有斜条纹的矩形块指示阅读框架移动。(a)一个 0—1 不对称外显子插入一个相位—0 内含子；(b)一个 0—0 对称外显子插入一个相位—1 内含子；(c)一个 0—0 对称外显子插入一个相位—0 内含子。(a)和(b)中的插入为不成功的插入。

(图 6—10)。外显子则根据其两侧的内含子而组合归类。例如，图 6—10b 中处在中间的外显子，其 5′端与相位—0 内含子相邻，其 3′端与相位—1 内含子相邻，因而说它是 0—1 类型。两端由同样相位的内含子包围的外显子称为对称的外显子(symmetrical exon)，否则即为非对称的(asymmetrical)。例如，图 6—10a 中处在中间的外显子是对称的。在 9 种可能的外显子类型中，3 种是对称的(0—0，1—1 和 2—2)，6 种是非对称的。

只有对称的外显子才能被插入内含子中。例如，图 6—11a 中，一个 0—1 外显子插入一个相位—0 内含子，结果造成后面所有外显子的阅读框架移动。而且，对称外显子的插入也是有限制的；一个 0—0 外显子只能插入相位为 0 的内含子，类似地，一个 1—1 外显子只能插入相位为 1 的内含子，一个 2—2 外显子也只能插入相位为 2 的内含子，以图 6—11b 的情况为例，一个 0—0 外显子插入了一个相位为 1 的内含子，结果造成被插入外显子和在它 3′端侧所有外显子的阅读框架移动。图 6—11c 则显示，一个 0—0 外显子插入一个相位为 0 的内含子之中将不会引起阅读框架移动。

6.9　产生新功能的变通途径

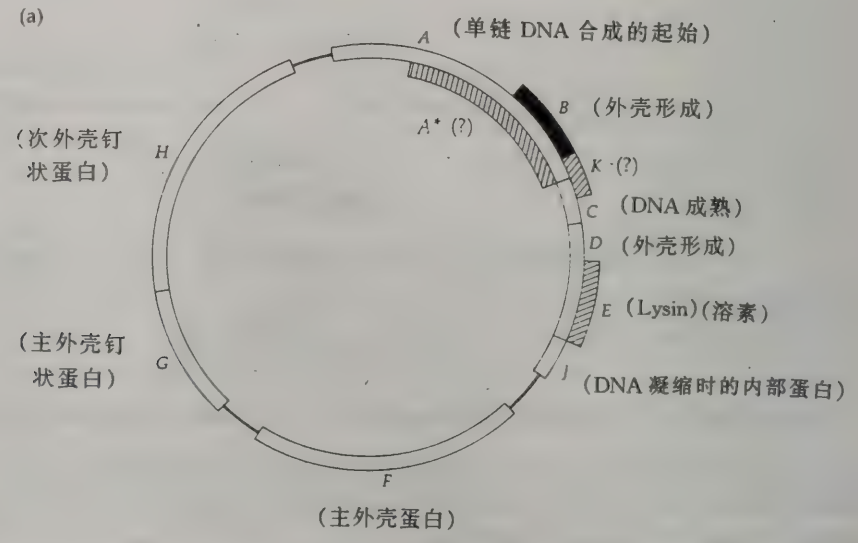
除了基因重复和外显子混匀以外，还有许多别的产生新基因或新多肽的机制。以下将考虑 3 种这样的机制。

重叠基因

已经发现，一个 DNA 片段能通过用不同阅读框架来为一个以上的基因编码。这一现象在病毒、细胞器和细菌中普遍存在。图 6－12a 展示了一个单链 DNA 噬菌体 ΦX174 的遗传图。其中已观察到几个重叠基因。例如，基因 B 整个地被包含在基因 A 之内，而基因 K 在 5′端与基因 A 重叠，在 3′端与基因 C 重叠。后一情况的更详细分析于图 6－12b 给出。

重叠基因也可通过应用一个 DNA 序列的两条互补链而产生。例如，人线粒体基因组中，确定 $tRNA^{Ile}$ 和 $tRNA^{Gln}$ 的基因分别位于不同的链上，并且它们之间有一个 3－核苷酸重叠，前者中读成 5′－CTA－3′，后者中读成 5′－TAG－3′(Anderson 等，1981)。

问题是，在进化期间重叠基因可能是怎样产生的呢？为了回答这一问题，我们注意到，开放阅读框架在整个基因组中大量存在。因此，相当长度的潜在编码区存在于已有基因的不同阅读框架中或互补链上，这是完全可能的。因为 64 种可能的密码子中只有 3 种是终止密码子，所以，即使一个随机的



(b)

```
K 蛋白：     M S R K I I L I K Q E     L L L L V Y E L
A 蛋白：...S D E S K N Y L D K A G I T T A C L R I
C 蛋白：
DNA 序列  ：  TCTGATGAGTCGAAAAATTATCTTGATAAAGCAGGAATTACTACTGCTTGTTTACGAATT

K 蛋白：   N R S G L L A E N E K I R P I L A Q L E
A 蛋白：   K S K W T A G G K *
C 蛋白：                   M R K F D L S L R S S R
DNA 序列  ：AAATCGAAGTGGACTGCTGGCGGAAAATGAGAAAATTCGACCTATCCTTGCGCAGCTCGAG

K 蛋白：     L L L C D L S P S T N D S V K N *
A 蛋白：
C 蛋白：   S S Y F A T F R H Q L T I L S K T D A L...
DNA 序列  ：AGCTCTTACTTTGCGACCTTTCGCCATCAACTAACGATTCTGTCAAAAACTGACGCGTTG
```

图 6－12　ΦX174 噬菌体单链环状 DNA 图。注意，为 B 蛋白质编码的基因(黑色)被完全包含在为 A 蛋白质编码的基因内，而基因 K 则与两个基因 A 和 C 重叠。自 Kornberg(1982)修改而成。(b)显示出与基因 A 的 5′部分和基因 C 的 3′部分重叠的 K 基因的顺序。星号表示终止密码子。(关于氨基酸的单字母缩写见表 1－1)。

DNA 顺序也可能含有成百个核苷酸长度的开放阅读框架。如果碰巧这样一个阅读框架中含有一个起始密码子和一个转录起始位点，或者通过突变产生了这些位点，那么，一条额外的 mRNA 就将会被转录出来，并随后被翻译成一个新蛋白质。这一新产物是否有有利功能那就是另一回事了，但如果它确有，则这种性状就可能会在群体中固定。

我们也注意到，为重叠基因编码的 DNA 区段上的进化速率，预期要低于只用一种阅读框架的类似 DNA 序列。其原因是，在重叠基因中非简并位点的比例要高于非重叠基因中的同类比例，这就大大降低了同义突变在总突变中的比例(Miyata 和 Yasunaga，1978)。

变通性的拼接

原始 RNA 转录产物的变通性拼接，可能会造成从同一个 DNA 片段产生不同多肽产物的结果。在这种情况下，外显子和内含子间的界限就不再是绝对的了，而是有赖于所涉及的 mRNA。许多 RNA 变通性加工的例子已在多细胞生物中被发现。

变通性拼接常被用作生长调节的手段。在涉及果蝇 *D. melanogaster* 的性别决定过程的几个基因中，曾看到一种非常有趣的情形。至少有 3 个基因：性致死基因(*Sxl*)、转化基因(*tra*)和倍性基因(*dsx*)，在雄性和雌性中是以不同的方式进行拼接的(图 6－13)。在 *dsx* 的情况中，该基因有 6 个外显子；外显子 1、2、3 和 4 用于雌性，而外显子 1、2、3、5 和 6 则用于雄性。在 *Sxl* 和 *tra* 的情况下，雄性中变通性拼接的产物含有成熟前终止密码子，因而是无功能的。例如，*Sxl* 的外显子 3 中含有一个框架内的终止密码子，但雌性的 mRNA 却不含这种外显子。

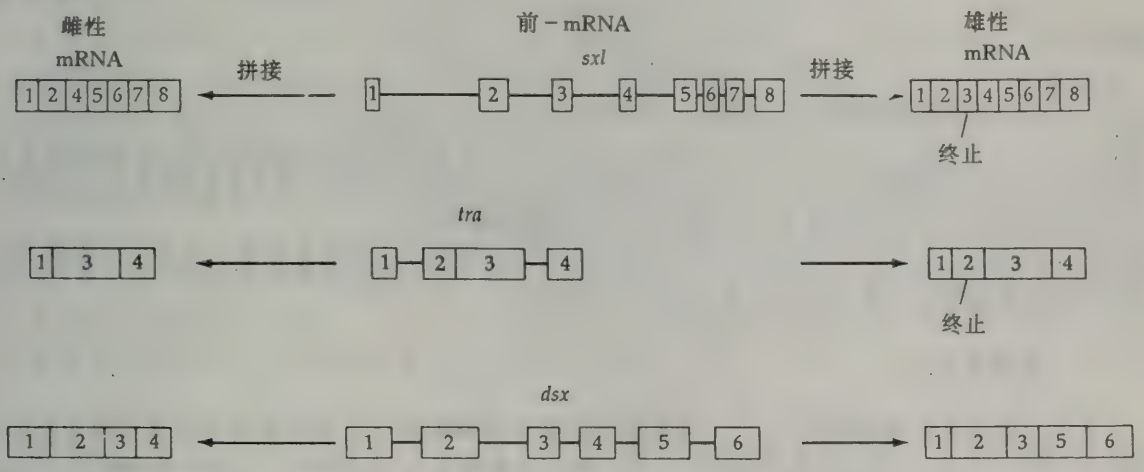


图 6－13　果蝇 *D. melanogaster* 雌性(左)和雄性(右)中，性致死基因(*Sxl*)，转化基因(*tra*)和倍性基因(*dsx*)的拼接模式。“停止”指示一个终止密码子，它截断成熟 mRNA 的编码区，从而造成产物无功能。自 Baker(1989)。

变通性拼接的一个特例可用由内含子编码的蛋白质的例子来加以说明(Perlman 和 Butow，1989)。在这类例子中，该内含子含有一个开放阅读框架，它为功能完全不同于其两侧外显子所编码的蛋白质的某种蛋白质全部或其一部分编码。在有些情况下，这类开放阅读框架是其上游外显子的延伸物，例如，酵母线粒体基因 *cox* I 中的内含子 a14α(图 6－14a)。在另一些情况下，内含子不仅包括一个游离态的编码蛋白质的基因，而且还含有关于转录起始和终止的必要信号(图 6－14b)。内含子 a14α 十分有趣，因为它为一种被称为成熟酶的酶蛋白编码。成熟酶对这个内含子从它的前体 mRNA 上准确地自我拼接去除是必要的。这种成熟酶在 DNA 重组中还起着内切核酸酶的作用。

要出现变通性拼接的进化，就需要重新产生一个变通性拼接的联结位点。因为拼接信号通常长为 5－10 个核苷酸，所以通过突变这类位点以一种可以查觉的频率产生是有可能的。事实上，从文献中知道已有许多这样的例子。例如，图 6－15 所示的例子中，甘氨酸密码子中的一次同义替换即把某一编码区变成了拼接部位。在图 6－15 关于 β^+-地中海贫血症的病理学查证的例子中，新的拼接位点通常比老的拼接位点更强(即，该突变发生后合成的 mRNA 大多数为已发生改变的那种类型)。这样的突变显然具有有害的后果，预期绝不会在群体中固定。然而，如果新产生的拼接位点要弱得多，则大多

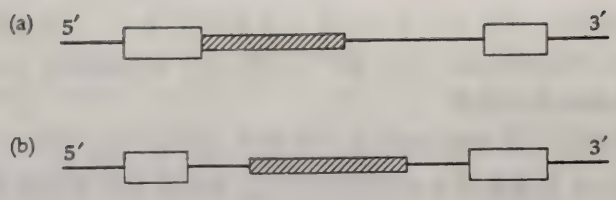


图 6－14　内含子为蛋白质编码的例子：(a)一个开放阅读框架(画有斜条纹)，它是上游外显子(空心矩形块)的一个延伸物(例如，酶母线粒体基因 *cox* Ⅰ中的内含子 al4α)；(b)一个游离存在的开放阅读框架，其转录起始和终止信号位于内含子之中(例如，噬菌体 T4 中 *sum* γ 基因的内含子)。资料取自 Perlman 和 Butow(1989)。

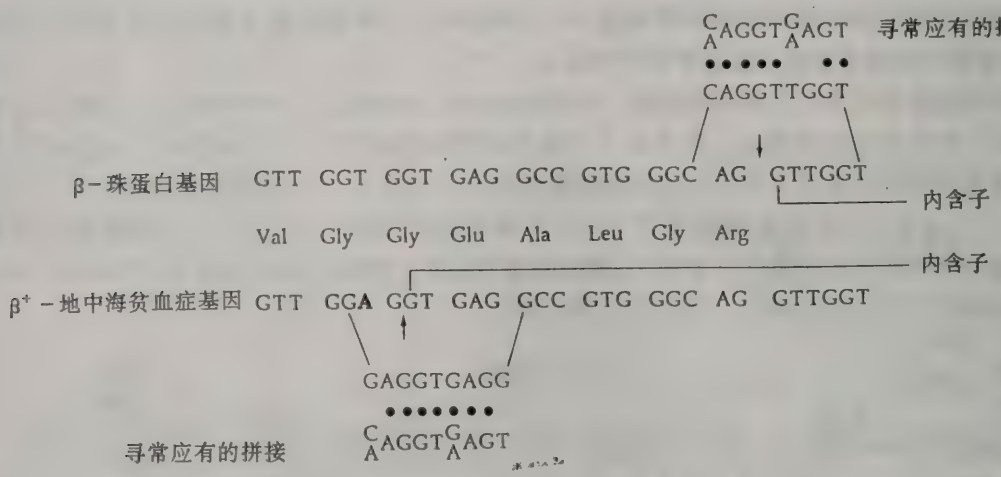


图 6－15　来自正常个体和患 β -地中海贫血症的病人的 β- 珠蛋白基因中，外显子 1 和内含子 I 间区域的核苷酸顺序。发生了突变的核苷酸以黑体字表示。箭头指示拼接位点。每个拼接部位都与寻常应有的拼接部位进行顺序比较，圆点表示这些拼接部位与寻常应有的拼接部位间有相同的核苷酸。

数 mRNA 将是原始类型，且只有少量的新型 mRNA 产生出来。这样的改变将不会抹杀旧功能，也为产生一种具有新功能的有用的蛋白质创造了一个机会。

基因分享

从产生新功能的观点看，当一个基因产物不作任何氨基酸顺序方面的改变而用来行使另外的功能时，一种令人极感兴趣的情形就出现了。这一现象曾被命名为“基因分享”(“gene sharing”)(Piatigorsky 等，1988)。基因分享的意思是，一个基因在没有重复也没有失去原始功能的情况下获得了并保持着第二种功能。不过，基因分享可能会要求在组织特异性或发育时序性的调节系统方面发生一点变化。

基因分享最初在晶状体中发现，这里晶状体是构成眼睛的水晶体赖以维持透明和适当的光线折射的物质。最初的发现是，来自鸟类和鳄类的∈晶状体在氨基酸顺序上与乳酸盐脱氢酶 B(LDH－B4，见第 84 页)等同，且具有同样的 LDH 活性(Wistow 等，1987)。后来的工作表明，这“两种”蛋白质事实上是同一种蛋白质，而且是由同样的基因来编码的(Hendriks 等，1988)。第二种晶状体 δ 存在于所有鸟类和爬行类之中，也已被证明在顺序方面与另一种酶等同。这种酶即精氨琥珀酸裂解酶，它催化将精氨琥珀酸转化成精氨酸的反应。这两个蛋白质好象也是由同样的基因编码的(Piatigorsky 等，1988)。类似地，七鳃鳗、真骨鱼类、爬行类和鸟类中的 τ－晶状体，已被证明与 α－烯醇酶等同并由同样的基因编码。α－烯醇酶是糖酵解中的一种酶，将 2－磷酸甘油酸转化成磷酸烯醇式丙酮酸(Piatigorsky 和 Wistow，1989)。于是，δ－、∈－和 τ－晶状体的例子说明，一个未经重复的基因能通过基因分享而获得额外的功能。另一方面，α、β 和 γ 晶状体则是另一类蛋白质的经典例子，这类蛋白质通过基因重复，以及其后由祖先基因分化成为不同蛋白质编码的基因而进化(例如，热震惊基因，为暴露于过热环境后才表达的蛋白质编码)。

基因分享可能是相当普通的现象。事实上，在以上例子中，那些酶和晶状体自身就可能有两种以上的功能。例如，τ—晶状体/α—烯醇酶也能象一个热震惊蛋白质那样起作用。显然基因分享增添了基因组的简洁性，即使在真核生物中简洁性看来并没有很高的优越性(第八章)。还要注意，在晶状体基因分享的例子中，同一个多肽既起着酶的作用又有结构蛋白质的功能，这就搅乱了酶和非酶、或结构蛋白质之间的传统界限。

6.10 多基因家族的协同进化

从十九世纪七十年代中期到十九世纪八十年代中期，关于DNA重退火和DNA杂交的研究大量涌现，目的在于探明真核生物基因组的结构和组织。这些研究揭示，较高等生物的基因组是由高度重复顺序、中度重复顺序和单拷贝顺序等3类序列所构成的(第八章)。它们还揭示出一个有趣的进化现象，即：一个重复顺序家族的成员在一个物种内相互间一般是非常相似的，而来自不同物种的该家族的成员，即使这些物种间亲缘关系很近，相互间也可能是很不一样的。这一现象最先被布郎等(Brown等，1972)查觉，他们是在比较来自非洲蟾蜍 *Xenopus laevis* 和 *X. borealis* 的核糖体DNA时发现的，后一种蟾蜍那时曾被误认为是 *X. mulleri*。

在 *Xenopus*(爪蟾属)和大多数别的脊椎动物中，确定18S和28S核糖体RNA的基因以成百的拷贝数存在着，且以一列或几个串联的列的形式排列着。每一重复单位由一个转录片段和一个不转录片段构成(图6—16)。转录片段产生一个45S RNA前体，该前体经酶切割而被划分成有功能的18S和28S核糖体RNA。这种转录重复片段通过不转录间隔片段(NTS)而相互隔开。

在对 *X. laevis* 和 *X. borealis* 间的核糖体RNA基因的比较中，布朗等(Brown等，1972)发现，虽然这两个物种的18S和28S基因极为相似，但这两个物种间的NTS区域却大不相同。相比之下，在每一个体内以及一个物种的不同个体间，该NTS区域则是非常相似的。于是，看起来好象是这样一种情况：虽然NTS区域在不同物种间迅速分岐，但它们在每一物种中却是一起进化的。布朗等(Brown等，1972)的结论是，一定有一种"校正"机制在起作用，以使某一突变从一个间隔顺序传向邻近的间隔顺序，其速度快于在这些顺序中出现新的变化。他们称这种在一个个体内显现的现象为水平进化(horizontal evolution)，以用来与垂直进化相对照。垂直进化是指某一突变在一个繁殖群体中的传播。后来，有人提出了并发进化(coincidental evolution)或协同进化(concerted evolution)等等术语。后一个术语由齐默尔等(Zimmer等，1980)提出，是目前文献中最通用的术语。

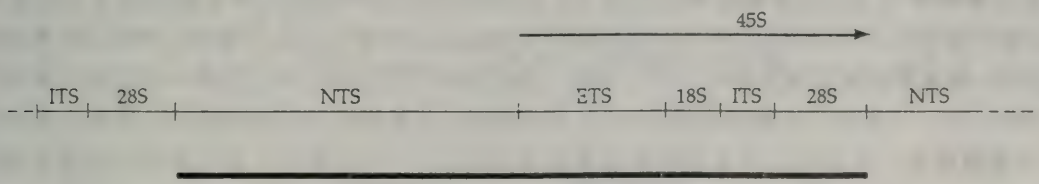


图6—16 脊椎动物rRNA基因的一个典型重复单位的图解表示。黑色杆线表示重复单位，箭头指示转录单位。ETS，外转录间隔；ITS，内转录间隔；NTS，不转录间隔。自Arnheim(1983)。

随着限制酶分析和DNA顺序测定等技术的出现，已有大量资料证明多基因家族中协同进化的普遍性(见Ohta，1980；Dover，1982；Arnheim，1983等综述)。图6—17展示了一个来自人和黑猩猩核糖体基因的限制酶分析的例子。人类中，每一重复单位在28S基因3′端的NTS区中有一个 *Hpa* I 位点，而在黑猩猩和其他大型猿类中则缺少这一位点。该 *Hpa* I 位点很有可能是在人—猿分枝之后而在人谱系中产生的，并且最终在每一个人的重复中固定了。NTS区中其他限制性位点也同样展示出物种特异的同源性。

协同进化本质上意味着，一个基因家族的某个成员并不是与该家族的其他成员毫不相干地进化着的。通过其成员间的遗传相互作用，一个多基因家族是以协同的方式，象一个整体一样地一起进化着。

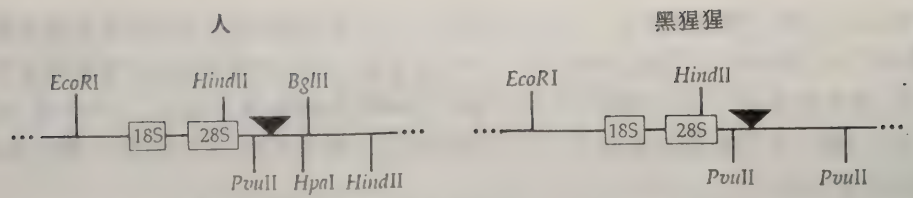


图 6－17　人和黑猩猩 18S 和 28S 核糖体基因中的限制性位点。所用限制酶为 *EcoR* Ⅰ，*Hind* Ⅱ，*Pvu* Ⅱ，*Bgl* Ⅲ，和 *Hpa* Ⅰ。基因上面标出的限制位点在物种中是多态的。基因下面的那些位点则是单态的。倒三角形表示 NTS 中长度上的多态性。自 Arnheim(1983)修改而成。

协同进化的机制

不等价交换(unequal crossing-over)和基因转变(gene conversion)(图 6－18)近来被认为是造成协

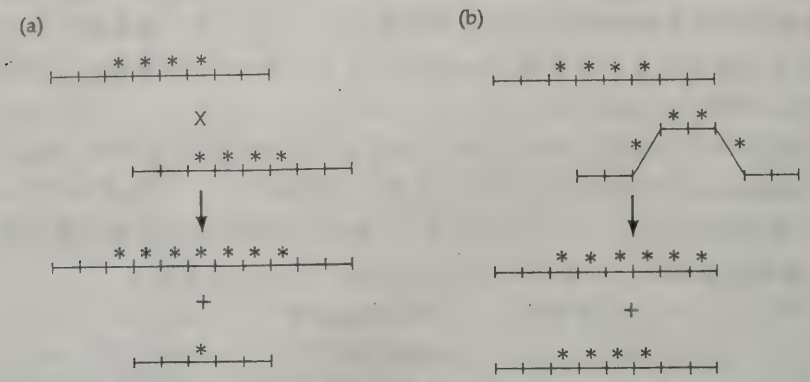


图 6－18　(a)不等价交换模型和(b)基因转变模型。不等价交换的结果是，两条子染色体都出现了重复数上的改变和两种重复类型(其中一种用星号标出)的频率上的改变，后者是与亲本频率(50%)相比较而言的。另一方面，基因转变则仅在其中一条子染色体中改变两类重复的频率，而且对两条染色体都不改变它们的总重复数。自 Arnheim (1983)修改而成。

同进化的两个最重要的机制。不等价交换可以发生在生殖细胞减数分裂时某一染色体的两条姐妹染色单位体间，也可以发生在有丝分裂时的两同源染色体间。一个交互重组的过程就是在一条染色单体或染色体中产生某一顺序的重复，而在另一条中则造成相应的缺失。图 6－18a 展示出的例子中，一次不等价交换事件导至一条子染色体上出现 3 个重复段的增幅，而在另一条上则出现 3 个重复段的缺失。这种不等价交换的结果是，两个子染色体都变得比它们的亲本染色体更为同源化。如果这种过程反复发生，则每种变异型重复在染色体上的数目将会随时间而波动，最后将有一种类型会在该家族中处于优势。图 6－19 是一个假想出来的例子，其中类型－4 基因通过反复多轮的不等价交换而传遍了某一个基因家族。不等价交换曾用数学方法详细研究过，并且已取得了相当程度的实验支持(见 Ohta，1980；Dover，1982；Li 等，1985a 等综述)。

基因转变是一种非交互重组的过程，在此过程中两个序列相互作用的方式为，其中一个被另一个转化(见 Lewin，1990)。从协同进化过程的观点看，基因转变中最重要的类型是非等位基因转变(即，位于不同基因座位的基因间的转变，而不是不同的等位基因形式间的转变)。图 6－18b 是一个非等位基因转变的例子，其中野生型重复中有两个转化成了突变型。结果，第一条子染色体变得比亲本染色体更为同源些，而在第二条子染色体中却未发生变化。理论研究表明，和不等价交换一样，基因转变也能产生协同进化(Ohta，1984；Nagylaki，1984)。基因转变曾经作为 γ-珠蛋白基因(Jeffreys，1979；Scott 等，1984)和另外许多基因中(见 Dover，1982)出现同源化的机制而提出来过。

作为一种协同进化的机制，基因转变看来有几个胜过不等价交换的优点。首先，不等价交换使一个家族中的重复基因的数目发生改变，故而有时可能会造成严重的份量不平衡。基因转变则相反，并不造成基因数目改变。其次，基因转变不仅能对串联的重复、而且能对分散的重复起着校正机制的作用。相比之下，不等价交换在所涉及的重复散布于非同源染色体上时就受到了限制。如果这些重复基

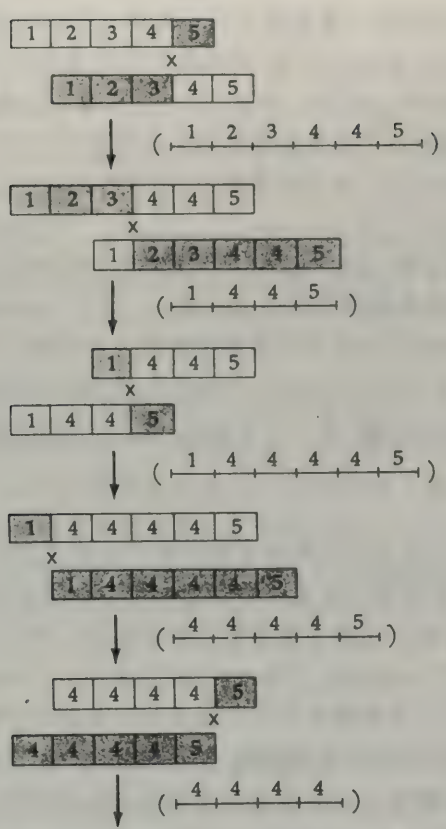


图 6—19　由不等价交换导至的协同进化。不等价交换事件反复循环使得每一染色体上重复基因变得越来越同源化。该过程还会影响每一染色体上重复顺序的数目。自 Ohta,(1980)。

因位于染色体的端粒部位(染色体臂的末端),象在人和猿的 rRNA 基因中的情况那样,或许它能对非同源染色体起有效作用;而若分散的重复位于染色体的中间,象在小鼠的 rRNA 基因中的情况那样,则它将受到极大限制(Arnheim,1983)。如果这些重复散布在一条染色体上,则不等价交换可能会导致重复段间基因的缺失或重复这样的结果。例如,图 6—20 表示两重复簇间不等价交换的一个假想

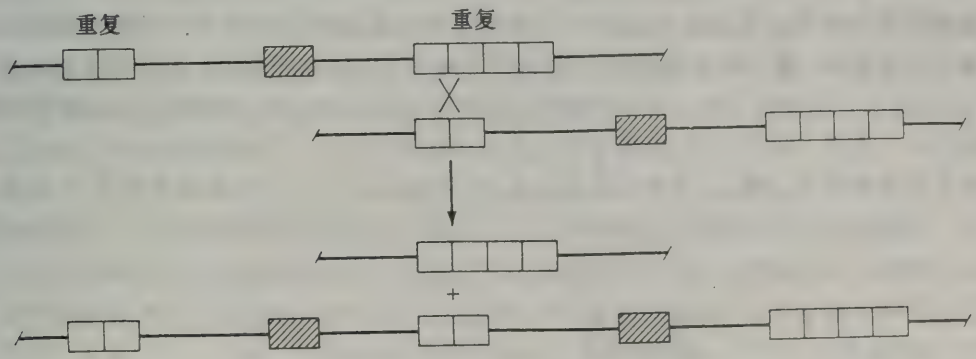


图 6—20　涉及散布重复(空心矩形块)的交换。画有斜条纹的矩形块表示一个唯一的基因。该交换事件后,这一基因在一条染色体上缺失,而在另一条染色体上则重复。

的例子,结果是一条染色体中某一唯一基因的缺失,而在另一条染色体中则相应地重复。其中一条染色体或者全部两条染色体,可能会对携带它们的生物带来有害影响。第三,基因转变可以有偏向性的方向。来自真菌的实验资料已经表明,基因转变方向上的倾斜是普遍的、而且常常是强烈的(Lamb 和 Helmi,1982),而理论研究业已证明,轻微的倾斜就可能对重复突变的固定概率有很大影响(Nagylaki 和 Petes,1982;Walsh,1985)。

在串联重复顺序的大家族中,不等价交换可能象基因转变一样是一个可接受的过程。首先,在这类家族中,重复的数目显然能极大地波动而不至引起显著的逆效应。这是从以下观察中得出的结论:

果蝇中确定 RNA 的基因的数目在同一物种的不同个体间，以及不同物种间变化幅度大(Ritossa 等，1966；Brown 和 Sugimoto，1973)。在人类中，已发现几个串联重复的家族，它们在拷贝数方面展现出异乎寻常的变化程度(Nakamura 等，1987)。其次，在一次基因转变事件中，通常只有一个小区域(异源双链区)涉及，而在不等价交换中，染色体间交换的重复数则可能非常大。显然，交换的重复数越大，协同进化的速率就越高(Ohta，1983)。在有些情形下，不等价交换的这一优点可大到足以与基因转变的优点相抗衡。

除不等价交换和基因转变之外，还有一些其他机制，象复制滑脱和转座(第一章和第七章)，也能造成某一家族中变异型基因的获得或丢失(Dover，1982)。最后，应注意到，协同进化不仅要求突变在该家族成员间的水平转移(同源化)，而且要求突变向群体中的所有个体传播(固定)。所以，我们还需要考虑随机遗传漂变的效应。多费(Dover，1982，1986)对在 DNA 转移和随机遗传漂变等各种机制联合作用下，多基因家族的协同进化过程，起了一个名称，即分子驱动(molecular drive)。

协同进化的进化论含意

协同进化使得某一变异型重复能传向所有基因家族的成员。这种水平地传播的能力有着深远的进化后果，因为这样一来，一个有利突变型重复就能替代所有其他重复而在该家族中固定。我们注意到，单单一个变异型所能给予生物的选择优势通常是很有限的。然而，如果该突变传给了许多甚至所有成员的话，则这种优势就会大大地加强。于是，通过协同进化，一个较小的选择优势可以变成较大的选择优势。在这方面，协同进化优于基因家族各个成员的独立进化(见 Arnheim，1983；Walsh，1985)。

阿恩海姆(Arnheim，1983) 曾对 RNA 多聚酶 I 的转录调控信号和 RNA 多聚酶 II 的转录调控信号的进化进行过比较。RNA 多聚酶 I 只转录 rRNA 基因，而 RNA 多聚酶 II 则转录所有为蛋白质编码的基因(第一章)。RNA 多聚酶 I 的转录调控信号的进化看来比 RNA 多聚酶 II 的该信号的进化要快得多。例如，在无细胞转录系统中，一种小鼠 rRNA 克隆不能在人的细胞提取物中转录，但来自差异极明显的物种的蛋白质编码基因的克隆却能够在异源系统中转录(例如，家蚕的基因在人的细胞提取物中，和哺乳类的基因在酵母的提取物中)。阿恩海姆(Arnheim，1983)认为，在关于 RNA 多聚酶 I 的转录单位的例子中，倾向于影响转录起始的那些有利突变，因协同进化所造成的后果可能已传播到整个 rRNA 多基因家族。与之不同的是，在关于 RNA 多聚酶 II 的转录单位的例子中，在任何一个基因中发生的影响转录起始的有利突变，预期将不会传遍所有基因，因为它们属于许多不同的家族。

关于新基因产生的传统观点是，先发生一次基因重复事件，然后该重复产生的两个基因之一逐渐分化而变成一个新基因。现已搞清，该过程可能不象以前所假定的那样简单。只要两个基因分岐的程度不是很大，则那个发生分化的拷贝就有可能通过不等价交换而被清除，或者通过基因转变而转化成保持原样的拷贝。在前一种情况下，它需要再发生一次重复以产生一个新的多余拷贝；而在后一种情况下则必须从头开始分化。所以，重复基因的岐化进程可能比传统上认为的要慢得多，为此缘故一个新基因从某一多余拷贝中产生的机会就减少了。另一方面，基因转变也可能会阻止一个多余的拷贝长时期地成为无功能状态，或者也许能有选择地使一个“死基因”(假基因)复活过来(Walsh，1987)。

我们已经习惯于假定，在一次基因重复之后，两个随之而来的基因将随时间单调地岐化着。在这样的假定下，我们前面已经证明，推测重复事件的时间是相当简单的。例如，人 β－和 δ－珠蛋白的蛋白质顺序相互间相似的程度比与兔 β1 或与小鼠 β 的主要和次要顺序的要高(Dayhoff，1972)。因此曾有过这样的推测：人的这两个基因是从约 4000 万年以前的一次重复事件中衍生而来的，这个时间远在哺乳类辐射(约 8000 万年以前)之后。考虑到重复基因能相互较正这样一个事实，则这一结论就可能是错误的。事实上，近来已经有人提出，β 和 δ 基因起源于发生在哺乳类辐射之前的一次重复(Hardison 和 Margot，1984)。该提议是根据这样的观察事实而作出的，兔假基因 ψβ2 的大内含子和 3′ 不翻译区与人的 δ 的相似程度比与兔 β1 的高，小鼠的假基因 βh3 相似于其 3′ 末端上的 δ。如果这一假说结果是正确的，则以上例子极好地说明了，基因一校正事件将会怎样部分或全部地抹擦掉重复基因间分岐进化的历史。在大的多基因家族中，基因一校正事件预期是频频发生的，在这种情况下，追踪家族成员间的进化关系将会更为困难。

从进化论的观点看，多基因家族的进化和分群体的进化之间存在着某种类比。我们可以把多基因家族中的每一种重复看成是分群体中的一种同类群。那么重复间信息的传递就等价于同类群间基因或个体的迁移。众所周知，迁移将减少两同类群间遗传差异的量，但会增加一个同类群中的遗传变异的量（例如等位基因的数目）。类似地，重复间的信息传递将会减少重复间的遗传差异但将增加某一基因座位上的遗传变异的量（Ohta，1983，1984；Nagylaki，1984）。小鼠主组织相容复合体中某些基因座位是高度多态的，事实上，已观察到某一基因座位上的等位基因数多达50个。所以，曾有人提出，这种程度较高的多态性是由基因转变所造成的（例如 Weiss 等，1983；但可参阅 Hughes 和 Nei，1989）。

习题

1、与某一随机选出的 DNA 片段的重复相比，外显子重复有什么优点？

2、在重叠基因中，简并位点的数目能大大地被削减。（a）如果下面的顺序仅按第1个阅读框架翻译，那么将有多少非简并位点？多少四重简并的位点？（b）如果除了第一个阅读框架外，该顺序也按第2个阅读框架翻译，那么将有多少非简并位点和四重简并位点？（c）如果3个阅读框架都进行翻译，那么该顺序中的非简并位点和四重简并位点将是多少？（3个阅读框架的起始点各用箭头标出）。

CATTCGTCTTTATTCGAAATCGCGTGGACAGCGGTGGATCTCTTTGCGCTGTGCAAAGCAGCGCTGGCGGTT

1
2
3

3、许多多聚体蛋白质是由重复基因编码的亚基所构成。有两种可能情况：（a）这些亚基可能全部来自一个基因座位，也可能来自不同的基因座位。在前一种情况下该蛋白质被称为是“同质型的”，而后一种情况下则是“异质型的”。假定该蛋白质象乳酸脱氢酶（LDH，见第84页原版）那样是一种四聚体酶，且其所有亚基由两个基因座位编码。那么，能产生多少种不同的同功酶？（b）这样的蛋白质常常是异质型的（即，常常是由不同基因座位产生的亚基所构成的）。例如，哺乳类成体的血红蛋白是一个由两条α链和两条β链构成的四聚体。假定在某一种哺乳动物中有3个α样基因座位和2个β样基因座位。那么，如果每一个四聚体是由两个α样基因座位产生的亚基和两个β样基因座位产生的亚基所构成的话，则能产生多少不同的异质型四聚体？

4、在大鼠和小鼠的基因组中有两个为胰岛素编码的基因（前胰岛素原Ⅰ和Ⅱ），而在除啮齿类以外的哺乳动物中则只有一个胰岛素基因。前胰岛素原Ⅰ基因被认为是通过一个所谓“反录转座”（第七章）的过程而产生的。前胰岛素原Ⅰ和Ⅱ这两个基因都在5′不翻译区中含有一个小内含子（长为118个核苷酸）。大鼠和小鼠中的内含子对子间，核苷酸差异数如下：

	内含子		
内含子	小鼠Ⅰ	小鼠Ⅱ	大鼠Ⅰ
小鼠Ⅱ	21		
大鼠Ⅰ	15	25	
大鼠Ⅱ	16	24	18

（a）该矩阵中的数字指示，小鼠前胰岛素原Ⅱ（即小鼠Ⅱ）基因中的内含子比其他基因中的相应内含子进化快，为什么？（b）假定核苷酸替换的速率恒定，并且假定小鼠和大鼠在1500万年前发生分歧，请用小鼠Ⅰ和大鼠Ⅱ但排除小鼠Ⅱ，来估计前胰岛素原Ⅰ和Ⅱ间的分歧时间。

5、在以下顺序的内含子（虚线）中插入一个0－0对称外显子。那么对阅读框架而言将会发生什么？如果插入一个2－2对称外显子会发生什么呢？如果插入一个不对称外显子又会发生什么？

5′——CAT TCG TCT TTA TTC GAA ATC GCG－－－TGG ACA GCG GTG AAT CTC TTT GAC GCT GTG——3′

6、为什么一个串联重复的家族能比一个散布重复的家族更容易经历协同进化？请解释。

后继阅读文献

Cold Spring Harbor Symposium on Quantitative Biology. 1987. *Evolution of Catalytic Function*. Vol. 52. Cold Spring Harbor Laboratory. Cold Spring Habor, NY.

Dayhoff. M. O. 1972. *Atlas of Protein Sequence and Structure*. Vol. 52. National Biomedical Research Foundation, Silver Spring, MD.

Dover, G. A. and R. B. Flavell. (eds) . 1982. *Genome Evolution*. Academic Press, New York.

Li, W. — H. 1983. Evolution of duplicate gene and pseudogene, pp 14—37. In M. Nei and R. K. Koehn (eds.) . *Evolution of Genes and Protein*. Sinauer Associates, Sunderland MA.

Ohno . S. 1970. *Evolution by Gene Puplication*. Springer-Verlag, Berlin.

Ohta. T. 1980. *Evolution and Variation of Multigene Families*. Springer-Verlage, Berlin.

7 由转座造成的进化

基因组曾经被认为是相当静止的实体，在其中可以给基因指定一个界限明确的基因座位。因此，基因被想象成在漫长的进化时期里一直停留在它们正好所处的染色体位置上。当芭芭拉·麦克林托克(Barbara McClintock)在十九世纪四十年代发现玉米中的某些遗传成份能从一个基因组座位"跳到"另一个上，不时地改变着结构基因的表达时，基因组的这种静态图景开始崩溃。然而，这种凝固静止的图景在科学思想界中已积习难改，以至于人们花了近40年的时间才认识到麦克林托克这一开创性发现的意义。今天我们认识到，基因组的结构组织比以前所想象的要更富于流动性、且更易于发生进化变化。在这一章里，我们将描述大量有助于遗传物质从一个基因组处位向另一个处位运动的可转座因子，并且讨论这类因子对进化过程可能会产生的影响。

7.1 转座与反录转座

转座(transposition)的定义是，遗传物质从一个染色体位置向另一个位置的运动。具有能改变其基因组位置这种内在潜能的 DNA 序列，被称为易动因子(mobile elements)或可转座因子(transposable elements)。根据可转座因子是否被复制来区分，则有两种类型的转座。在保守型(conservative)转座中，该因子从一个位点移到另一个位点(图7－1a)，供体位点上发生了什么则不清楚。有一种模型提出，供体 DNA 的末端不是相互连结的，因而转座后剩余下来的分子就瓦解了。然而，如果细胞含有该供体序列的重复，则该供体 DNA 种即可避免从细胞谱系中丢失。在这种情况下，虽然消耗了一个拷贝，但另一个却幸存了下来，由此而产生的谱系将在原始位点上有一个因子，而在新位点上有第2个因子(Berg 等，1984)。另一种模型提出，该双链断裂是由宿主的修复系统所修复的。

在复制(型)(replicative)或重复(型)(duplicative)转座中，可转座因子被拷贝，并且其中一个拷贝留在原位点处，另一个则插入一个新位点中(图7－1b)。所以，复制型转座的特征是可转座因子的拷贝数的增加。有些可转座因子只采用一种类型的转座，而另一些则对保守型和复制型两种方式全采用。

在以上转座类型中，遗传信息由 DNA 所携带。已知遗传信息也能通过 RNA 而转座。在这种模式中，DNA 被转录成 RNA，然后 RNA 再反转录成 cDNA(图7－1C)。为了将这两种模式加以区别，我们把由 RNA 中介的模式称为反录转座(retroposition)。转座与反录转座都曾既在真核生物中又在原核生物中发现(见 Weiner 等，1986；Temin，1989)。与由 DNA 中介的转座相比，反录转座总是复制型的，因为被转座的是该因子的反转录拷贝，而不是该因子本身。

当一个可转座因子插入一个宿主基因组中时，在该插入位点处的一小段宿主 DNA(通常为4－12bp)被复制(图7－1)。这种经复制的重复有同一取向，因而被称为顺向重复(direct repeats)。这是转座和反录转座的特征标记。

有些可转座因子能转座到所有细胞中去；而另一些则是高度特异的。例如，果蝇 *D. melanogaster* 的 *P* 因子通常仅在生殖细胞中是易动的。转座的接受位点的基因组位置在不同可转座因子中也表现出有变异。有些因子对某一特别的基因组位置表现出不寻常的偏爱。例如，*IS4* 总是精确地将自己掺入到大肠杆菌的半乳糖苷酶操纵子中的同一点上，所以，每个细菌只能含有一个拷贝的 *IS4*(Klaer 等，1981)。另外一些，象噬菌体 *Mu*，则能随机地转座到几乎任何基因组位置上。许多可转座因子则表现出中等程度的基因组位置偏向性。例如，*E. coli* 的 *Tn*10转座子，其40%被发现位于 *lac Z* 基因中，成为基因组的精细结构部分；*P* 因子对 *X* 染色体有亲合性，且偏向于插入从5′端靠近基因编码区的序列中，而不是编码区之中。有些可转座因子对某种特别的核苷酸组成展示出有更高的亲合性。例如，*IS1*

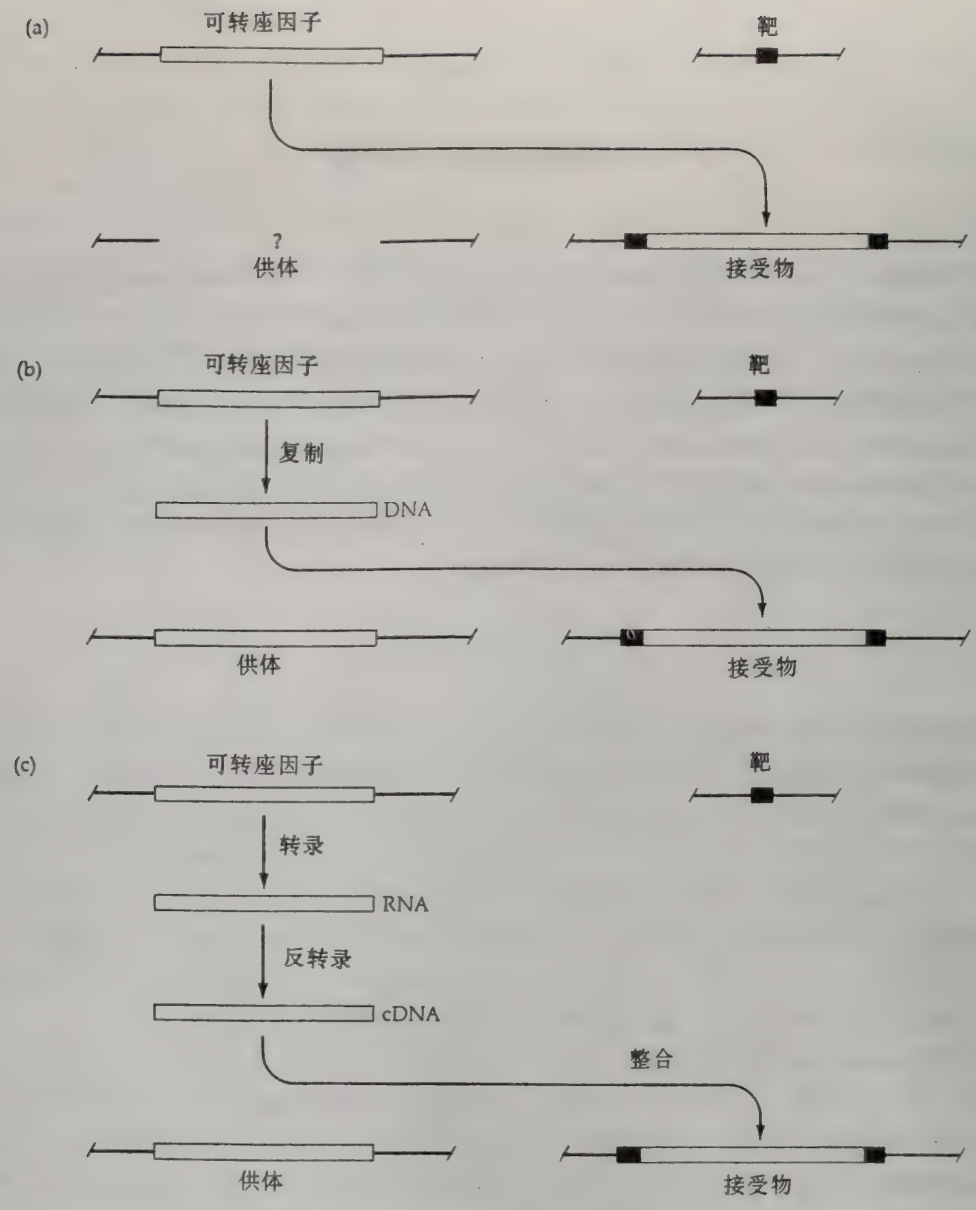


图7－1 (a)保守型转座。该因子从供体位点转座到靶位点。供体位点处发生了什么则不清楚。该供体分子可能会瓦解，这对于具有一个以上染色体拷贝的细菌而言是可以忍受的。另一种可能性是，该双链的断裂处被宿主的修复系统所修复。(b)复制型转座。该因子被复制，并且有一个拷贝插入一个靶位点处，而另一个拷贝则保留在供体位点上。关于保守型和复制型转座的更详细解释可参阅 Lewin (1990)。(c)反录转座。该因子转录成 RNA，然后 RNA 再反转录成 DNA。该 DNA 拷贝插入宿主基因组。作为反录转座的一个例子可见图7－4。注意，转座与反录转座都会在新插入成份的两端各产生一小段重复(黑矩形块)。

偏爱富含 AT 的插入位点(Devos 等，1979)即是如此。

7.2 可转座因子

可转座因子，根据其转座模式和其所含基因的数目与类型，可分成三类：插入序列、转座子和反录因子。

插入序列

插入序列(insertion sequences)是最简单的可转座因子。它们除具有为转座所必需的部分外不带

任何遗传信息。插入序列在长度上通常为700－2500bp，已在细菌、噬菌体、质粒和玉米中发现。细菌的插入序列由前缀 *IS* 加上后随的类型号数来表示。一个来自肠道细菌 *E. coli* 和 *Shigella dysinteria* 的插入序列，*IS*1的结构，如图7－2a 所示。*IS*1在长度上约为770个核苷酸，包括两个颠倒的非等同末端重复，每个各23bp。它含有两个阅读框架，*InsA* 和 *InsB*，它们为一种或两种形式的转座酶(transposase)编码。转座酶是一种催化可转座因子向插入位点插入的酶。在 *E. coli* 中有几十种不同类型的插入序列，从自然界分离出的品系，其大多数的基因组中所含每种序列的数是可变的(Sawyer 等，1987)。

转座子

转座子(transposons)是易动因子，通常长为2500－7000bp 左右，大多数在基因组中作为散在重复顺序的家族而存在。它们与插入序列的区别是还带有所谓外源性基因(exogenous genes)，即一些为除与转座有关的功能之外的功能蛋白质编码的基因。(注意，在有些文献中命名方式较混乱，有时转座子这个术语被用来指所有可转座因子，包括插入序列、反录转座子等。)在细菌中，转座子用前缀 *Tn* 和后随的类型号数来表示。有些细菌转座子是复合转座子(complex tremsposons)或混合转座子(composite transposons)，这样命名是因为，有两个完整而独立的可转座插入序列以任意一种排向从两个侧面夹拥着一个或多个外源性基因(图7－2b)。有趣的是，在复合转座子情况下，不仅整个转座子能作为一个整体转座，而且两侧的一个或两个插入序列还能独立地转座。因为转座的功能是由插入序列编码的，所以，复合转座子通常不含有一种独立的转座酶基因。

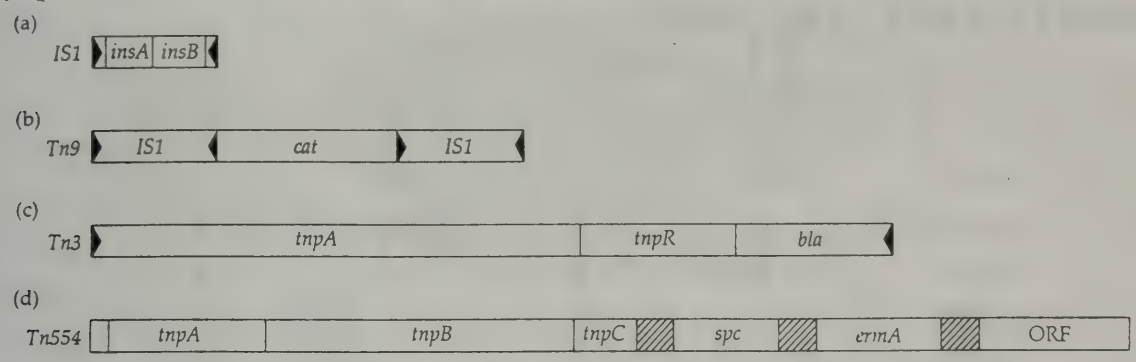


图7－2　细菌中的四种可转座因子的模式图。黑三角形表示颠倒重复。(a)来自 *E. coli* 和 *Shigella dysenteria* 的插入序列 *IS*1两侧由长为23bp 的不完美颠倒重复所夹拥。(b)来自 *E. coli* 的复合转座子 *Tn*9，含有两拷贝 *IS*1，分别位于 *cat* 基因两侧。*cat* 基因为具有氯霉素抗性的蛋白质编码。(c)来自 *E. coli* 的转座子 *Tn*3，能产生链霉素抗性，含有3个基因，其中两个(*tnpR* 和 *bla*)在一条链中转录，第三个(*tnpA*)则在另一条中转录。*Tn*3两侧都有长为38bp 的完美颠倒重复。(d)来自 *Staphilococcus aureus* 的 *Tn*554缺少末端重复，含有5个基因和1个开放阅读框架(ORF)。其中3个基因(*tnpA*，*tnpB* 和 *tnpC*)为转座酶编码，且作为一个单位而被转录。*spc* 和 *ermA* 基因则分别提供对壮观霉素和红霉素的抗性。*spc* 基因为依赖 S－腺苷甲硫氨酸的甲基化酶编码，是在与别的基因不同的链中转录的。ORF 被大量转录，但是否被翻译则尚不得而知。画有斜条纹的矩形块不含开放阅读框架。

其他的细菌转座子，以及许多真核生物的转座子，其两侧只有较短的不同取向的重复顺序(图7－2c)，并且不含插入序列。然而，并非所有转座子都是在结构上对称的。有些具有不对称的末端，缺少颠倒的或顺向的末端重复(图7－2d)。动物中有些转座子(例如果蝇中的 *P* 因子)的编码区被内含子所隔断(图7－3)。

图7－3　果蝇 *D. melanogaster* 的 *P* 因子的模式结构图。该因子两侧有长为31bp 的短颠倒重复，其编码区含有4个外显子(白矩形块)，由3个内含子(黑矩形块)所隔断。该因子长为2900bp 左右。

细菌中的转座子常常带有这样一些基因，它们能给予携带者以抗生素抗性(如 *Tn*554)、重金属抗性(如 *Tn*21)或抗热性(如 *Tn*1681)。质粒则可把这样一些转座子从一个细胞带到另一个细胞，结果，

抗性就能迅速传遍到整个暴露于这样一些环境因子中的细菌群体中。

有几种噬菌体事实上是细菌中的转座子，或转座性噬菌体(transposing bacteriophages)例如，噬菌体 *Mu* 就是一个非常大的转座子(～38000bp)，它不仅为那些调控其转座的酶编码，而且还为大量构成其 DNA 包装所必需的结构蛋白质编码。

许多类型的转座子在动物、植物和真菌的基因组中广泛分布。例如果蝇 *D. melanogaster* 就含有50—100种不同类型的转座子，而且都是多重拷贝的(Rubin，1983)。

反录因子

反录因子(retroelements)是含有一个为反转录酶编码的基因的 DNA 序列或 RNA 序列，反转录酶则催化以 RNA 为模板的 DNA 的合成。由此而产生的 DNA 分子称互补(的)DNA(complementary DNA. cDNA)。这些的确转座的反录因子是通过反录转座过程而转座的。有一些不同类型的反录因子，我们采用由特明(Temin，1989)提出的分类，并将它们列于表7—1。

反录病毒(retroviruses) 是一些结构上类似于转座子的 RNA 病毒。虽然它们是所有反录因子中最复杂的，但我们还是最先讨论它们，因为反录转座概念起源于对反录病毒生命周期(图7—4)的发现。反录病毒的颗粒称为病毒粒子(virion)，它侵入某一宿主细胞以后，其基因组 RNA 即被反转录成病毒 DNA。该 DNA 能整合到宿主基因组中，变成一个前病毒(provirus)。接下来，前病毒 DNA 转录成 RNA，它们既可作为合成病毒蛋白质的 mRNA，又可作为病毒的基因组，并被包装到具有感染性的病毒粒子中。病毒粒子一旦形成，周期即可再次开始。

表7—1 反录因子和反录序列的分类

因 子	反转录酶	转座	LTR[a] 是否存在	病毒粒子
反录子	是	否	否	否
反转录子	是	是	否	否
反录转座子	是	是	是	否
反录病毒	是	是	是	是
拟反录病毒	是	否	是	是
反录序列	否	否	否	否

自 Temin(1989)。

a，LTR，长末端重复。

反录病毒至少具有3个基因：*gag*，*pol* 和 *env*(图7—5)。这些基因分别为几种内部蛋白质、几种酶(包括一种反转录酶)和一种被膜蛋白编码。许多反录病毒还有别的基因：例如艾滋病毒至少有6个额外基因。反录病毒的编码区两侧有长末端重复(long terminal repeats，LTRs)。LTR 含有与转录(在前病毒阶段)有关的启动子和与反转录(在病毒阶段)有关的启动子。

反转录子(retroposons)和反录转座子(retrotransposons)是不构成病毒粒子的可转座因子，所以，它们与反录病毒不同，不能独立地穿越细胞而转座。它们间的相互区别是，是否存在末端重复顺序(LTR)(表7—1)。注意，有些作者是将反转录子和反录转座子当作同义词来用的。果蝇中的 *copia* 因子代表一种典型的反录转座子；它的两端都具有 LTR，并且还含有一段长的开放阅读框架，其中有类似于反录病毒的 *pol* 基因的区域。图7—5b 所示的是另一个反录转座子的例子，即粘性霉菌 *Dictyostelium discoideum* 中的 *DIRS*—1因子。*D. discoideum* 平均含有40个左右的完整 *DIRS*—1拷贝，以及大约200—300个的 *DIRS*—1片段。有趣的是，*DIRS*—1有一种将自己插入别的 *DIRS*—1序列中，从而抵消它们的作用的倾向，这或许是粘性霉菌基因组中存在许多有缺陷的 *DIRS*—1片段的原因(Cappello 等，1984)。*DIRS*—1基因的转录受发育阶段诱导，也受热震惊的诱导。

与反录转座子不同，反转录子不含 LTR。果蝇 *D. melanogaster* 中的 *G3A* 因子是一个反转录子(图7—5c)。这种反转录子含有两个 ORF。ORF—1含有一个与反录病毒的 *pol* 基因类似的区域，而 ORF—2则含有7个由非常短的间隔顺序隔开的外显子。

反录子(retrons)是最简单的反录因子(图7—5d)。它们曾在某些细菌基因组中发现(Inouye 等，

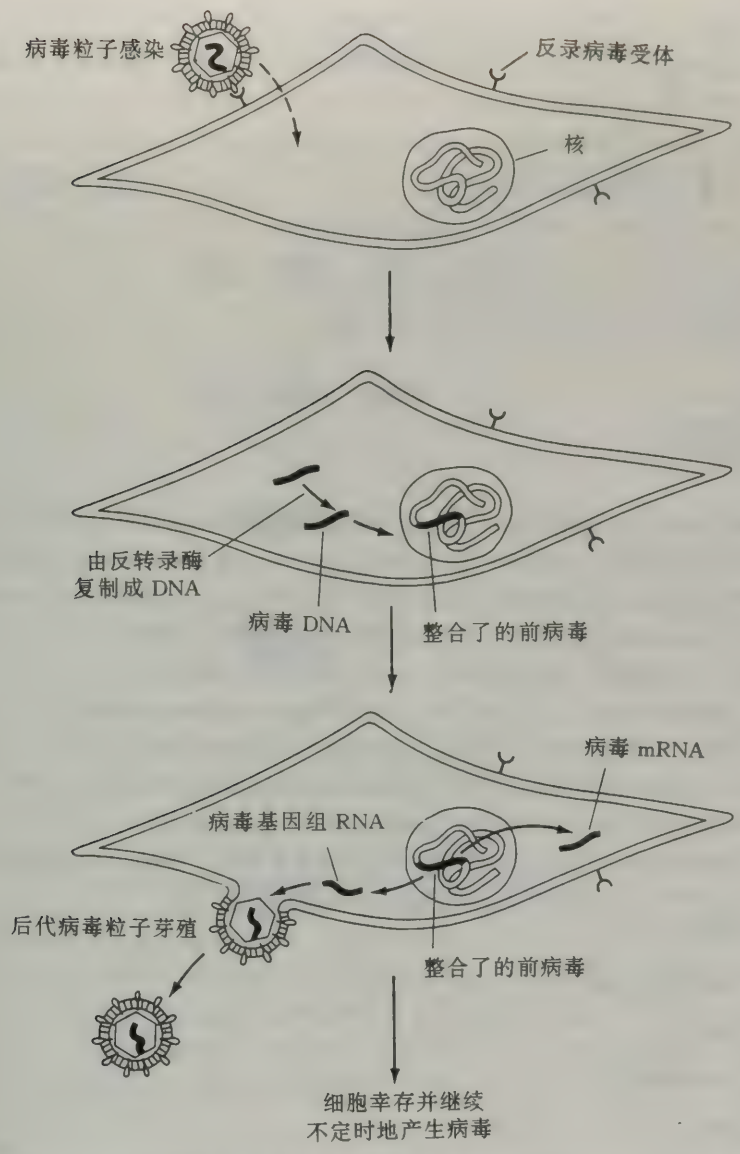


图7－4　一个反录病毒生命周期的概貌。反录病毒的病毒粒子附着到细胞表面的一个受体上。基因组 RNA 注入细胞质中，在那里通过反转录酶的作用而被反转录。cDNA 穿入细胞核，并且可能会整合到宿主细胞的基因组之中。整合进去的前病毒被转录成，(1)用于合成病毒蛋白的 mRNA，和(2)基因组 RNA。该基因组 RNA 与结构蛋白质和酶性蛋白质装配成有感染性的病毒粒子，以出芽的形式穿出细胞膜。自 Watson 等(1987)。

1989；Lampson 等，1989)，也曾在植物 *Oenothera berteriana* 的线粒体基因组中发现(Schuster 和 Brennicke. 1987)，它们的开放阅读框架中有与其他反转录酶基因类似的顺序。然而，反录子不切离，因而是构成基因组整体的必要部分。与前病毒不同，反录子没有 LTR，也不能构成病毒粒子。

拟反录病毒(pararetroviruses)，如乙型肝炎病毒，是结构上类似于反录病毒，但已失去了将自己插入宿主基因组中的能力的一类病毒。由于这个缘故它们是不够作为可转座因子的资格的，虽然它们显然与反录病毒有着共同的进化起源。

所有反录因子的反转录酶都有某些氨基酸等同性，这一事实表明了这些因子有一种共同的进化起源。由于反录子与反录病毒的复杂结构相反具有简单性，又由于细菌的古老性，所以，特明(Temin，1989)认为，进化的途径是从反录子到反转录子、到反录转座子、到反录病毒、到拟反录病毒(图7－6)。当然，某些现存的反录转座子有可能是从反录病毒衍生而来，而不是从它周围的另一通路衍生而来。

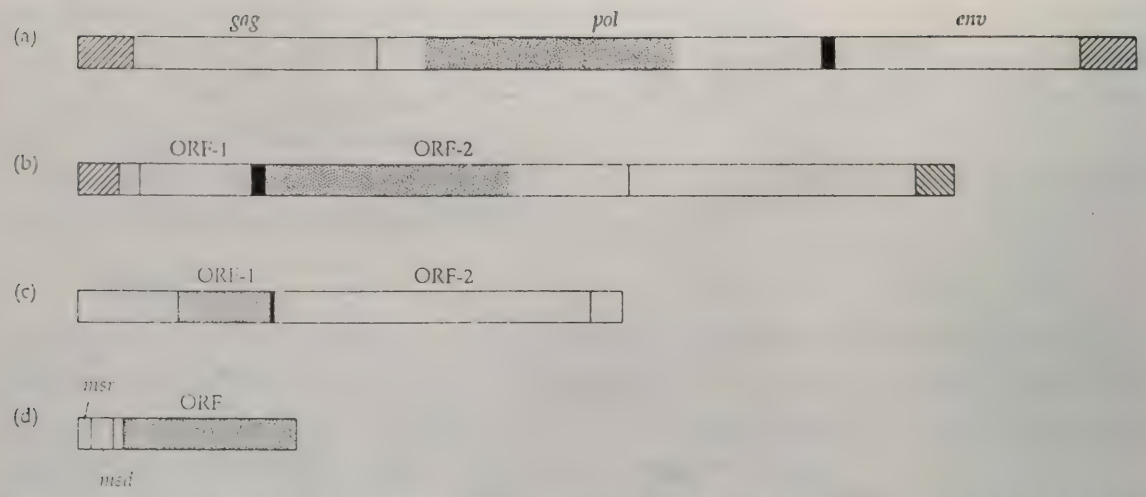


图7－5　反录因子的模式结构。打点的区域表示为反转录酶编码的区域；相邻基因间的重叠区则用黑矩形块表示；画有斜条纹的区域表示长末端重复(LTR)。ORF，即开放阅读框架。(a)猫白血病反录病毒。编码区的两侧有两个不等同的LTR，各长482和472bp。该编码区为两个多重蛋白前体编码。由 *gag-pol* 区编码的多重蛋白前体将分裂成两个多重蛋白，分别对应于 *gag* 和 *pol*。*gag* 多重蛋白产生4种内部病毒蛋白质，分别用p15，p12，p27，p10表示。*pol* 多重蛋白则分裂成3种酶，一种蛋白酶，一种反转录酶，和一种内切核酸酶/整合酶。由 *env* 编码的多重蛋白前体分裂成两种被膜蛋白、分别用p70和p15表示。(b)粘性霉菌(*Dictyostelium discoideum*)的反录转座子 *DIRS*－1。颠倒的LTR长为200－350bp。ORF－2含有一个顺序与反录病毒的 *pol* 基因类似的区域；(c)果蝇 *D. melanogaster* 的反转录子 *G3A*。ORF－1含有一个顺序与反录病毒的pol基因类似的区域。注意LTR缺乏这一特征。(d)来自粘球菌 *Myxococcus xanthus* 的反录子。其中，*msr* 基因被转录成RNA；而 *msd* 基因则从互补的链上转录，然后再经由该反录子的ORF编码的反转录酶反录成DNA。反录产生的两个分子接着通过2′、5′-磷酸二酯键而相互连结，从而形成被称为复拷贝单链DNA(msDNA)的分枝形分子。注意这里没有LTR。

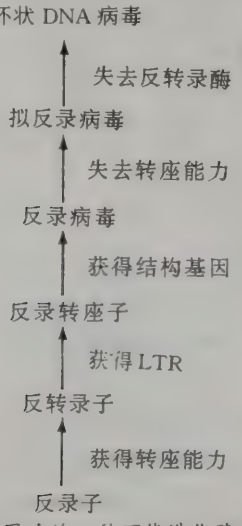


图7－6　反录因子的一种可能进化路线的模式图。

7.3　反录序列

反录序列(retrosequences 或 retrotranscripts)，是通过RNA的反转录而得到、接着整合到基因组中、但缺乏产生反转录酶能力的基因组的序列(表7－1)。产生反录序列的模板通常是某一基因的RNA转录本。有些作者把反录序列称为“非病毒的反转录子”(例如 Weiner 等，1986)。一种产生反录序列的过程如图7--7所示。如果某一基因不在任何种系细胞中转录，那么，产生反录序列就需要RNA跨越细胞障碍。这可以通过以下方式实现，即RNA分子被包进某一反录病毒的病毒粒子中，然后传送到种系细胞，并在那里被反转录(Linial，1987)。这一过程被命名为反录转染(retrofection)。

由于反录序列起源于RNA序列，所以它们带有一些进行过RNA加工的标记，因而又称为加工后序列(processed sequences)。反录序列的特征包括：(1)缺少内含子，(2)与基因的转录区域有精确一致的边界，(3)在3′端有poly-A延伸物，(4)在两端都有短的顺向重复，这指示可能已涉及转座，(5)各种转录后的修饰，如短核苷酸延伸物的加入或去除，以及(6)序列所在染色体上的位置已不同于转录产生RNA的原始基因的基因座位。

存在有两种类型的反录序列：加工后基因(或反录基因)和加工后假基因(或反录假基因)

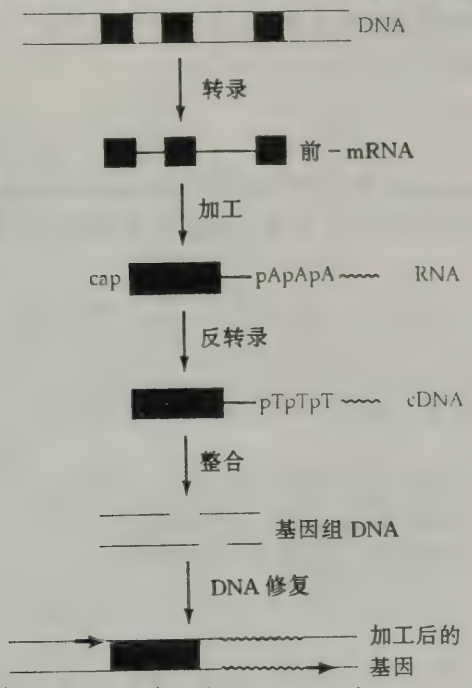


图7－7　加工后反录序列的产生。黑块状体表示外显子，波纹线表示mRNA中的多聚A尾巴和cDNA中的互补多聚T。DNA被转录成前体mRNA，然后被加工成mRNA。mRNA再被反转录成cDNA，cDNA则整合进基因组DNA中。裂缝被修复，这样在插入的反录序列两段侧就产生了两段顺向的短重复(黑水平箭头)。如果cDNA从合成RNA的细胞转插入到一个不同细胞的基因组中，则产生反录序列的过程就需要mRNA包装进一个反录病毒粒子中，并且转移到那个靶细胞里。这样的过程称反录转染。

反录基因

加工后基因(processed gene)或反录基因(retrogene)是一种有功能的反录序列，它产生的蛋白质与产生该反录基因的原基因所产生的蛋白质等同或接近等同。有几个原因可以说明为什么一个反转录的基因保留其功能的可能性很小。第一，反转录过程是很不精确的，以至RNA模板与cDNA之间可能会出现许多差异(突变)。第二，除非加工后基因是从一个由RNA多聚酶Ⅲ转录的基因衍生而来，否则它通常不含有那些位于不转录区中的必需调控序列。第三，加工后基因有可能被插入到不适于其正确表达的基因组位置上。事实上，在绝大多数情况下，一个加工后基因就是一个“垂死者”。

令人惊奇的是，加工后的有功能基因已被发现，虽然它们看来是非常罕见的。人的磷酸甘油酸激酶(*PGK*)多家族由一个活性的X－连锁基因、一个加工后的X－连锁基因和一个额外的常染色体基因所组成。该X－连锁基因含11个外显子和10个内含子。另一方面，它的常染色体同源物却是不寻常的，它没有内含子并且它的3′端由一个poly-A尾巴的剩余物构成侧翼，这极有力地表明曾发生过涉及mRNA的反转录过程。有趣的是，该常染色体的*PGK*基因几乎独特地在睾丸中表达。于是，反转录而成的*PGK*基因不仅保留了完整的阅读框架和转录并产生有功能的多肽的能力，而且还获得了一种新的组织特异性(McCarrey和Thomas，1987)。鸡中的肌肉特异的钙调素(calmodulin)基因也是无内含子的，显然也是经由反转录酶中介的事件而产生的(Gruskin等，1987)。

大鼠和小鼠的前胰岛素原Ⅰ基因可能是半加工反录基因(semiprocessed retrogene)的一个代表性例子。该基因在5′不翻译区中有一个长119bp的内含子。相比之下，它的同源物前胰岛素原Ⅱ，含有与

它同样的小内含子，另外还有一个存在于c肽编码区中的较大(499bp)内含子。来自其他哺乳类，包括其他啮齿类在内的所有前胰岛素原基因也含有两个内含子。而且，前胰岛素原Ⅰ基因两侧由短重复夹拥，且在多聚腺苷化信号后有一段短多聚A(poly-A)区(Soares 等，1985)。这些特征表明，前胰岛素原Ⅰ基因可能是从被部分加工的前胰岛素原Ⅱ基因的前体mRNA衍生而来的半加工反录基因。事实上，根据这两个前胰岛素原基因间的比较，前胰岛素原Ⅰ看来是从一个偏离了正轨的前体mRNA转录本衍生而来的，该转录本起始于正常帽子部位上游的500个碱基对处，且只有第一个内含子从它里面切离了。正是因为该偏离了正轨的转录本含有未被正常转录的5′调控序列，反录基因在整合到一个新基因组位置后才保留了它的功能。

加工后假基因

加工后假基因(processed pseudogene)或反录假基因(retropseudogene)是一个失去其功能的反录序列。它带有一切有功能的反录序列的标记，但却有一些妨碍其表达的分子缺陷。一个有功能的基因

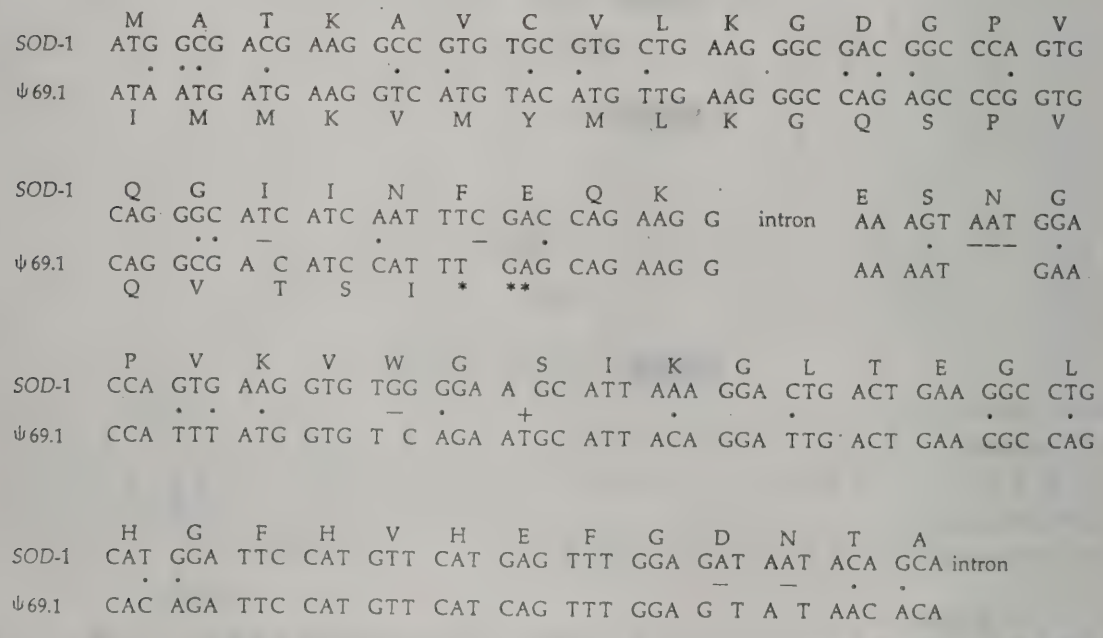


图7－8 人的Cu/Zn超氧化物歧化酶基因(*SOD*-1)的前两个外显子与一个加工后假基因(Ψ69.1)的同源部分间的比较。圆点表示替换，"－"表示缺失，"＋"号表示插入。注意内含子的缺乏和成熟前终止密码子(用星号指示出)。关于单字母氨基酸缩写可参阅表1－1。资料自 Danciger 等(1986)。

和一个加工后假基因间的比较如图7－8所示。许多加工后假基因在反录转染期间被截去端部；加工后mRNA的5′截尾特别普遍，但3′截尾也是曾有所闻的。5′端被截去可能发生在：(1)转录期间(例如从正常位置的下游起始转录)，(2)RNA加工期间(例如错误地拼接)，或(3)反转录期间(例如酶在反转录进行到RNA分子的3′端之前失效，该部位对应于cDNA的5′端)。

已经知道了从所有类型的RNA(例如mRNA，tRNA，rRNA，snRNA和7SL RNA)衍生而来的加工后假基因。转运RNA特别有趣，因为它们提供了一个最有说服力的证据，表明加工后假基因事实上是通过RNA的反转录而派生出来的。所有细胞核tRNA都在3′端有一个CCA序列(图7－9)。该序列不由确定该tRNA的基因编码，而是转录后经酶作用加上去的。对比之下，基因组的加工后tRNA假基因则常常在3′末端有CCA顺序。

加工后假基因已在动物、植物、甚至细菌中发现。不过，虽然加工后假基因在哺乳类中大量存在，但它们在其他生物，象鸡、两栖类和果蝇中却相对地较为稀有。表7－2列出了人类和啮齿类中的某些加工后假基因，这些假基因以及它们的功能基因的数目都是已知的，或已估出的。平均下来，在这些物种中加工后假基因的数目比功能基因的数目要多。事实上，在许多情况下加工后假基因的数目甚至有可能是被低估了的，因为古老的加工后假基因可能在顺序上与其亲本基因发生了较大程度的岐化，以

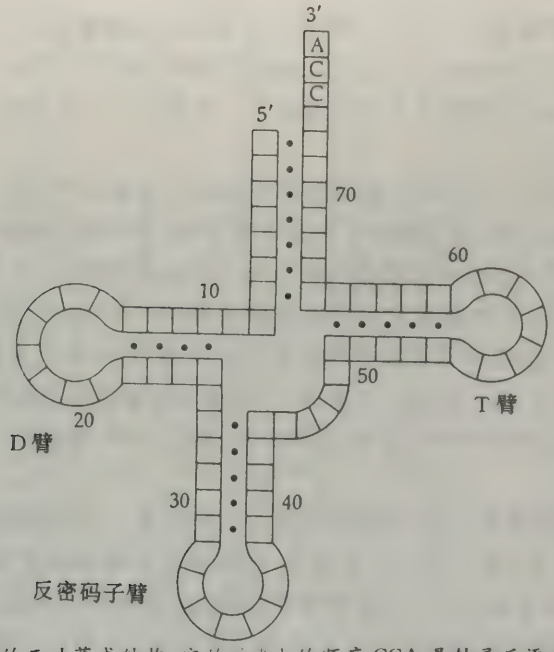


图7－9　一个tRNA分子的三叶草式结构。它的3′端上的顺序CCA是转录后添加到功能分子中的，但它常在tRNA假基因的基因组顺序中出现。

至用从其功能基因同源物制造出的分子探针已不再能将它们检出来了。

表7－2　反录假基因和其亲本功能基因的数目

物种	基因	基因的数目	反录假基因的数目
人	精氨琥珀酸合成酶	1	14
	β－肌动蛋白	1	～20
	β－微管蛋白	2	15－20
	Cu/Zn 超氧化物歧化酶	1	≥4
	细胞色素C	2	20－30
	二氢叶酸还原酶	1	～5
	非肌的原肌球蛋白	1	≥3
	甘油醛－3－磷酸脱氢酶	1	～25
	磷酸甘油酸激酶	2[a]	1
	核糖体蛋白L32	1	～20
	磷酸丙糖异构酶	1	5－6
小鼠	α－珠蛋白	2	1
	细胞角蛋白内A(cytokeratin endo A)	1	1
	甘油醛－3－磷酸脱氢酶	1	～200
	肌球蛋白轻链	1	1
	鸦片黑素皮质激素原	1	1
	核糖体蛋白L7	1－2	≥20
	核糖体蛋白L30	1	≥15
	核糖体蛋白L32	1	16－20
	肿瘤抗原P53	1	1
大鼠	α－微管蛋白	2	10－20
	细胞色素C	1	20－30

自 Weiner 等，(1986)。

a、其中一个是反录基因。

在有些情况下，加工后假基因的数目可以超过其有功能对应物数目达几个数量级。*Alu* 家族就是一个这样的例子，该家族之所以如此命名，是因为这种序列中含有一个 *Alu* 1内切核酸酶的特征性限制位点。*Alu* 序列长约为300bp，它们属于人类基因组中一个超过50万次的重复序列家族，以构成基因组的5－6%而引人注目。

乌卢和楚迪(Ullu 和 Tschudi，1984)发现，*Alu* 序列实际上是确定7SL RNA 的基因的加工后假基因。7SL RNA 在切除分泌蛋白质的信号顺序中是至关重要的。其活性基因受着严格限制，而且它的顺序在象人、爪蟾以及果蝇这样一些分歧程度很高的生物中都是保守的。人的 *Alu* 序列是从7SL 序列经过一系列步骤，包括一次重复，两次缺失和多次核苷酸替换(图7－10a)衍生而来的。大多数的人 *Alu* 序列有一个二聚体结构。人基因组还含有许多四聚体 *Alu* 序列，但至今只有几个单体 *Alu* 因子曾在人类中发现。相比之下，啮齿类的 *Alu* 等价物 *B*1家族，几乎是绝对地单体的(图7－10b)。布里滕等(Britten 等，1988)将第一个单体出现的时间定在哺乳类辐射之前的年代，而把产生二聚体的重复所处的年代定在灵长类谱系建立以后。

如果一个反录假基因保留着被转录的能力，则其后可能会有一个倾落过程跟随，藉此新的反录假基因即从现存反录假基因的 RNA 转录中产生出来。这种情况曾被认为在 *Alu* 家族中发生过(Bains，1986)。我们注意到，7SL 基因是被 RNA 多聚酶Ⅲ转录的，而该酶则并不需要转录区外的启动子。因此，某些 *Alu* 序列保留了完整的启动子并且连续不断地被转录，这是可以想象到的。不过，威拉德等(Willard 等，1987)和布里滕等(Britten 等，1988)仅认出了 *Alu* 序列的几个亚家族。布里滕等(Britten

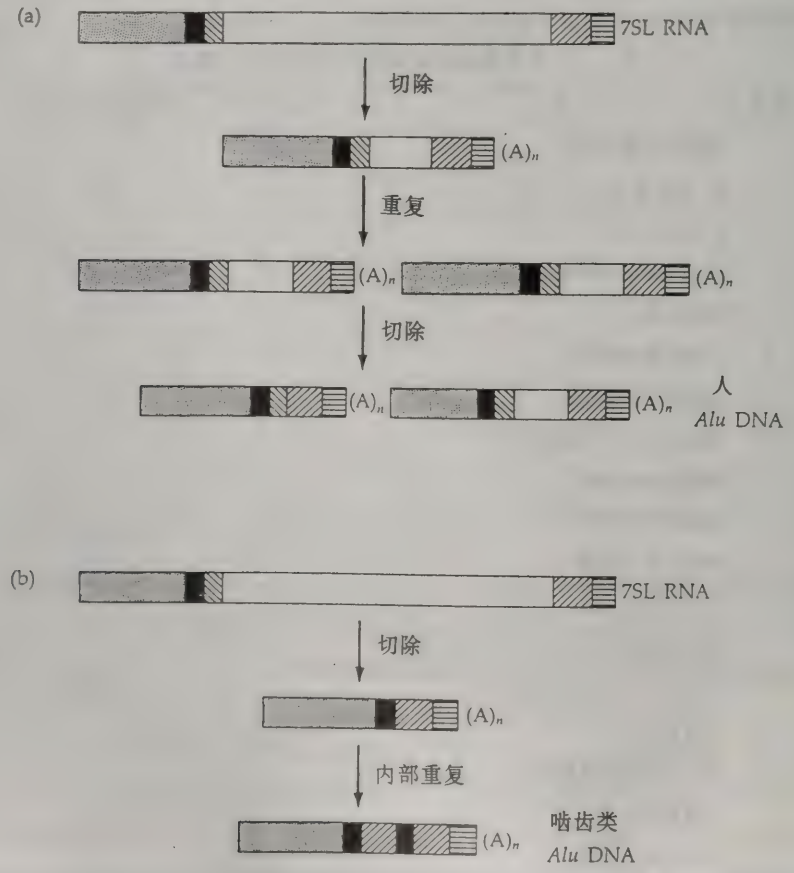


图7－10　*Alu* 序列的起源(a)在人和其他灵长类中(b)在啮齿类中7SL RNA 基因中的不同区域用不同的阴影图区别，以强调 *Alu* 序列中的缺失和重排。(A)n 表示 A 被重复 n 次。注意在(a)中的二聚体结构和在(b)中的单体结构。

等，1988)提出，这些亚家族是从4个源基因逐步衍生的(图7－11)。更近期衍生出来的亚家族，其与祖先7SL 序列在顺序上的分歧比古老一些亚家族的更大。所以结论是，*Alu* 序列不是直接从7SL 功能基因衍生而来，而是从少数源序列衍生而来的，这些源序列则是从7SL RNA 序列经过许多变化步骤而产生的。这四种源基因各在这一时刻或那一时刻起着 *Alu* 序列的主导源作用，并各被某一后代谱系所替代。固定的连续波虽不在突然的爆发后出现，但连续的亚家族却继续在基因组中长期共存(另见

Quentin，1988）。

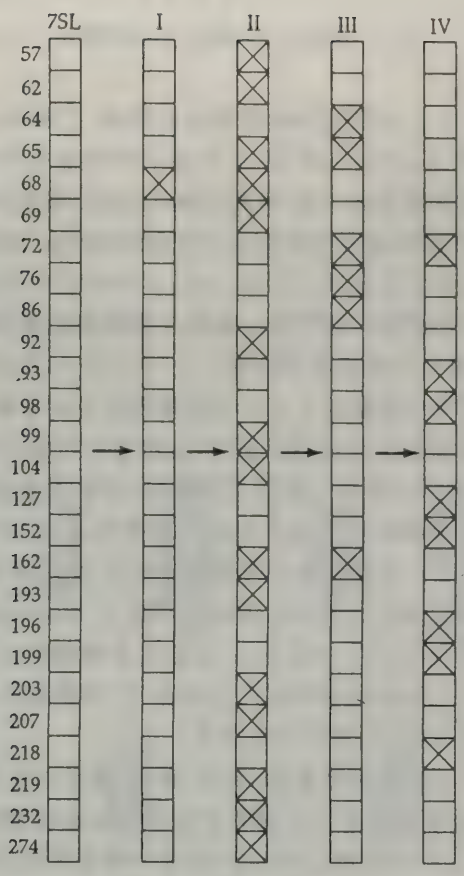


图7—11 *Alu* 序列的不同亚家族间在已判断出的位置（有阿拉伯数字）上的突变次序。罗马数字指示在不同时期起着 *Alu* 序列的主导源作用的连续的源基因。将每一个亚家族与其前面一个区别开来的替换用 X 表示。资料自 Britten 等（1988）。

由于反转录的泛有性，哺乳动物的基因组简直是处在反录序列的拷贝的困扰中。绝大多数的这类拷贝从它们整合到基因组中时起就是无功能的。而且，这样一些序列也不能通过基因转变而轻易地复苏，因为它们大多数位于距其亲本功能基因很远的染色体位置上（第六章）。一个有功能的基因座位泵出它自己的有缺陷拷贝、并将它们散布在整个基因组中的现象，可以与火山产生岩浆作类比，因此该过程曾被命名为**进化的维苏威模式**（Vesuvian mode of evolution）（P. Leder，在 Lewin，1981中被引用）。

加工后假基因的进化

正如前面所说的，一旦加工后假基因作为基因组中的一种染色体序列而存在后，它就是无功能的并不受一切选择的限制。由于缺少功能，所以假基因受两种进化过程的影响（Graur 等，1989b）。第一种涉及点突变的极迅速的积累。这种积累最终会抹去假基因与其功能同源物间的顺序相似性，因为后者的进化要缓慢得多。假基因的核苷酸组成将变得与其无功能的近邻越来越相象，以至最终它将“混入”其中。这一过程曾被称为**组成同化**（compositional assimilation）。

第二种进化过程的特征是，与其功能基因相比，假基因会变得越来越短。这种长度缩短（length abridgement）是由于缺失超过插入而造成的。曾有人估计，哺乳动物的加工后假基因在大约4亿年的时间里失去了约一半的 DNA。这一过程是如此地缓慢，以至在人的基因组中还含有那些在其非常远古的祖先中发现的假基因 DNA 的主要部分就是一个很好的例子。显然，这些古老的假基因时至今日几乎已经失去了与其从之发端的功能基因的一切相似性。

总而言之，看来加工后基因产生的速率比它们通过缺失而被抹去的速率要快得多。所以结论是，缩短的过程进行得太慢了，以至于无法抵消继不断的维苏威式的轰击之后出现的基因组大小上的增加（第八章；Graner 等，1989b）。

7.4 转座对宿主基因组的影响

转座和反录转座可能对基因组的大小和结构有着深远的影响。尤其是，可转座因子已被看成是“自私 DNA”(selfish DNA)的最好例子；这里自私 DNA 可能不带给宿主任何利益，却因为它比基因组的序列增殖迅速而能在基因组中传播(Doolittle 和 Sapienza，1980；Orgel and Crick，1980)。为此缘故，转座能极大地增加基因组的大小。这种效应将在第八章中探讨。这里，我们将只关心可转座因子影响基因进化和表达的方式。

第一，如以上所讲到的，细菌中的转座子常常带有能给予携带者以抗生素抗性或其他抗性的基因。于是，转座子也许能使该宿主物种在某种逆境中生存。

第二，一个基因的表达也许会因为可转座因子存在于该基因之中或其邻近而改变。情况最简单的是，可转座因子插入到一个为蛋白质编码基因的编码区中，这将极有可能改变阅读框架，因而可能会有激烈的表型效应。类似地，可转座因子的切离可能是不精确的，结果就会导至碱基的增加或缺失。然而，也有预料不到的效应。例如，一个可转座因子也许会含有调控因子，象启动子，这就会影响邻近基因的转录频率。事实上，反录病毒的 LTR 中常常含有很强的增强子，它们将对邻近基因的表达产生很大的影响。类似地，酿酒酵母 *Saccharomyces cerevisial* 中的 *Ty*(表示“transposon yeast”)因子已知能增加下游基因的表达。这在某些特殊环境中可能是有利的，虽然在大多数情况下，由这类变化造成的代谢不平衡极有可能是有害的。含有拼接给体或受体的可转座因子，即使被掺入到基因的非编码区，如某一内含子之中，也有可能会影响原初 RNA 转录本的加工。

第三，许多可转座因子能促成大型的基因组重排。倒位、易位，重复以及大段的缺失和插入可以经可转座因子的中介而发生。这些重排能以转座的直接后果发生(即，通过 DNA 片段从一个基因组位置向另一位置的运动)，并且能改变它们进行表达时的环境。作为转座的结果，若两个先前相互间很少相似性的序列现在能共享某一相似的可转座因子，以至它们间有可能进行不等价交换，则随后将会出现较间接的效应。图7－12示意了，由于复本 *Alu* 序列存在于低密度脂蛋白受体基因的外显子5侧面的内含子中，从而促成了一次不等价交换事件，而这样一次事件又是怎样导至一个缺少该外显子的突变

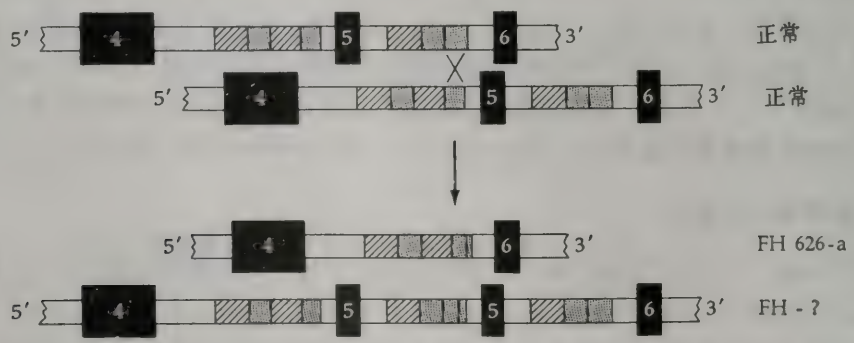


图7－12 低密度脂蛋白受体基因中的不等价交换。外显子用黑矩形块表示并加以编号。内含子中的 *Alu* 序列用画斜条纹(右臂)和加点(左臂)来表示；一个 *Alu* 序列可以是由两个臂(二聚体的)也可以是由一个臂(单体的)所构成的。设想的交换位置用 x 指示。重组后的缺失(观察到的)和插入(推测的)产物被描绘成 FH626－a 和 FH－?，位于箭头之下。自 Hobbs 等(1986)。

型基因的产生的(Hobbs 等，1986)。在低密度脂蛋白基因中，由同样机制造成的外显子14的缺失也已被观察到(Lehrman 等，1986)。对这些缺失呈纯合的病人在血液中有高水平的胆固醇(高胆固醇血症)。两个 *Alu* 因子间的重组还被证明是造成患腺苷脱氨酶缺陷症病人的腺苷脱氨酶基因缺失启动子和第一个外显子的原因(Markert 等，1988)。一般而言，对于所有含 *Alu* 重复序列的区域，基因组不稳定性都已得到证实(Calabretta 等，1982)。

第四，有证据表明，某些可转座因子可能会造成突变率增加。例如，含有可转座因子 *Tn*10的 *E. coli* 品系被发现插入速率升高了(Chao 等，1983)。在大多数情况下，这种性状对携带者来说将是有害

的。然而，在严峻的环境压力下，突变率提高也有可能是有利的，因为有些突变也许能更好地适应新的环境，而它们的携带者也将比非携带者有更高的适合度。

7.5 杂种劣势

果蝇中的杂种劣势(hybrid dysgenesis)是由一些相关的异常遗传性状造成的症候群，这些性状在某些相互作用的品系间的一个杂种类型中自发地诱生出来，但在相反方向交配的杂种中则通常不会出现(Sved，1976；Kidwell 和 Kidwel，1976)。杂种劣势曾引起分子生物学家和进化生物学家的浓厚兴趣，因为已经发现它是由可转座因子造成的，且它的主要特征是产生阻止品系间或群体间杂交的屏障，而这种屏障则曾被推测是物种形成的一个原因(见后)。

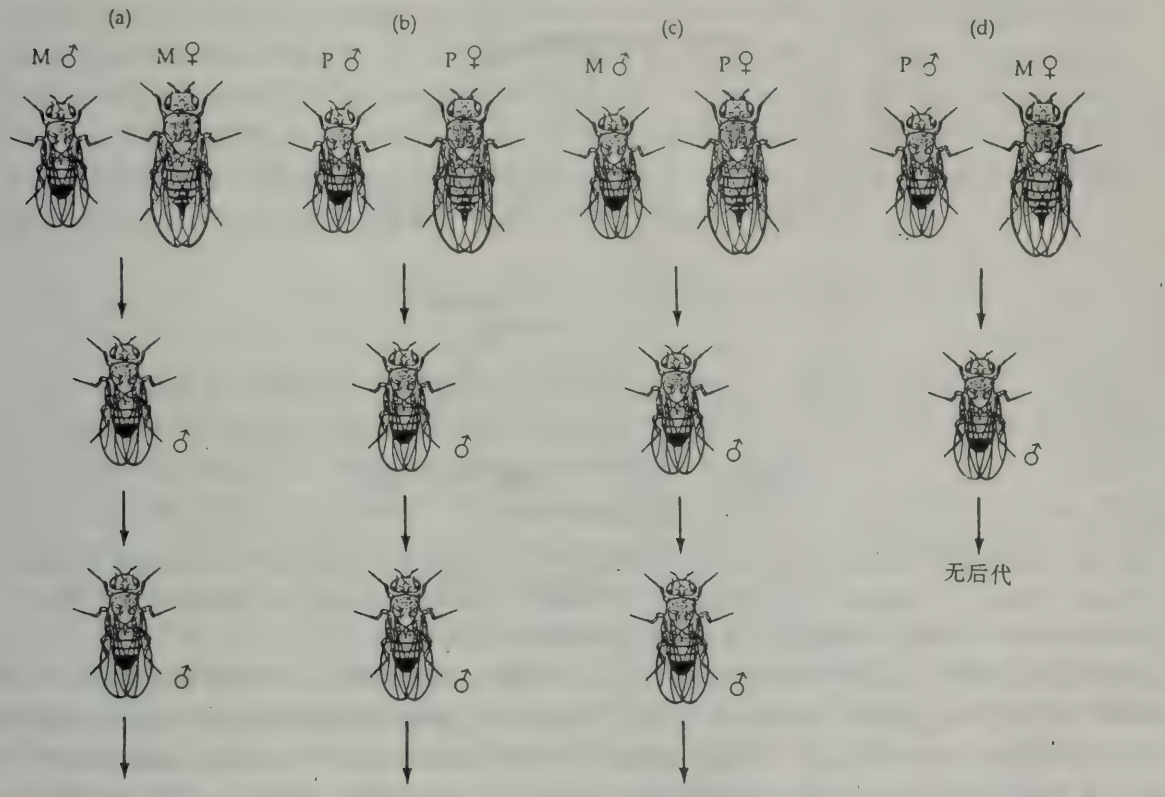


图7－13 果蝇中的杂种劣势。(a)*M* 品系内交配和(b)*P* 品系内交配产生正常后代。(c)正常后代也能在 *M* 雄体和 *P* 雌体的杂交中产生。(d)在 *P* 雄体和 *M* 雌体的杂交中，产生了劣势的后代，它们中许多是不育的。

果蝇中有几种杂种劣势系统，下面我们将仅对其中之一，*P*－*M* 系统进行探讨。杂种劣势的不对称性如图7－13所示。当一个 *P* 品系的雄体与一个 *M* 品系的雌体交配时，其后代是劣势的；而在相反方向的交配中，其后代则是正常的。*P*－*M* 系统的劣势性状包括：(1)某些个体在一定的条件下生殖腺发育不全，(2)在雄体中出现重组(在果蝇中重组通常被限制在雌体中，所以这是一种非自然的现象)，(3)染色体发生断裂，(4)偏离了孟德尔传递比率(即，相对于 *P* 染色体而言 *M* 染色体传给后代的倾向性要大一些)，和(5)出现高突变频率。

P－*M* 系统劣势的原因来自一个被称为 *P* 因子的可转座因子的家族(图7－3)。在 *P* 品系中，基因组中有30－50个 *P* 因子，但它们中许多都可能含有缺失。它们分布于所有染色体中。虽然在有些品系中，转座表现得有点偏爱X染色体。*M* 品系则不带 *P* 因子。杂种劣势系统的不对称性被认为是这样造成的，由于母性遗传，*P* 雌体×*M* 雄体交配的F1子代中存在着 *P* 因子编码的抑制物，而相反方向的交配 *M* 雌体×*P* 雄本，其F1子代中则缺少这种抑制物。*P* 因子可以在从母系的细胞质中得到缺乏该抑制物的种系细胞中转座，杂种劣势是与这种转座相关联的。在杂种劣势的背景下，细胞质中抑制物的存在与否定义了合子形成后的反应类型，因而被命名为细胞型(cytotype)。在抑制物存在下，*P* 因子的转座可能被完全或部分地抑制。正常情况下体细胞中不发生 *P* 因子的转座，因为体细胞中编码

区中的第3个内含子不象在种系细胞中那样会被切离。

基德韦尔(Kidwell,1983)观察到的一个有趣现象涉及带有 P 的品系的分布。在1950年以前收集的任何 *D. melanogaster* 果蝇品系中都未曾观察到 P 性状,而随后收集到的品系则表现出 P 的频率增

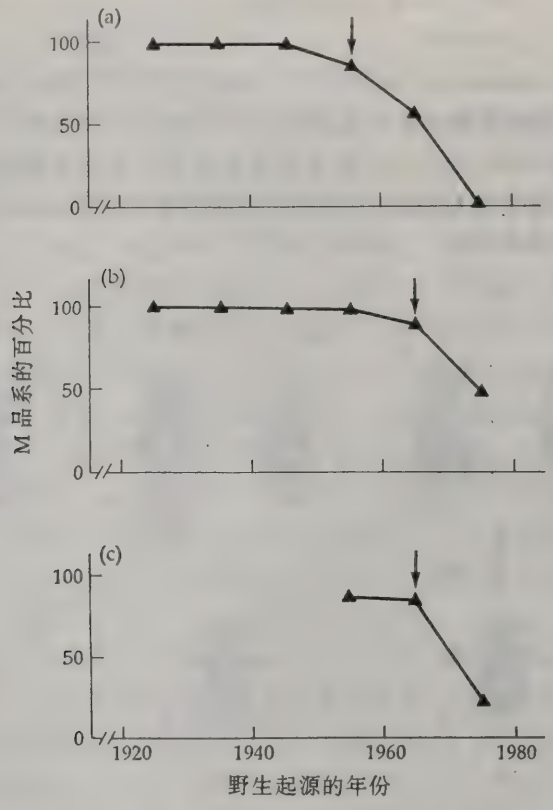


图7—14 自然群体中 M 品系(即无 P 因子的品系)频率变化,(a)在北美和南美的自然群体中,(b)在欧洲、非洲和中东的自然群体中,(c)在澳洲和远东的自然群体中。箭头指示 P 品系的第一次出现。注意,M 群体的比例下降最初出现在美洲大陆,只是后来才在其他大陆出现下降。自 Kidwell(1983)修改而成。

加和寿命降低(图7—14)。另一个劣势系统 I—R,已得到类似的观察结果。已有两种假说被提出来,以解释 P 因子的分布。恩格斯(Engels,1981b)提出,自然界中大多数品系是 P 型的,但它们在实验室群体中有失去该转座子的倾向。第二种假说则设想,P 转座子是近来才导入 *D. melanogaster* 群体的,随后 P 因子在以前的 M 群体中迅速传播(Kidwell,1979)。有几个原因可以说明为什么第二种假说看来要更合理一些。首先,带有 P 的实验室品系曾被监测了近15年,并未失去它们的 P 特性。其次,在 P 品系的分布上看来有一个地理渐变群,北美群体表现得比某些欧洲、非洲和亚洲品系更早地带有 P 特性。最后,现在已有证据表明,*D. melanogaster* 近来已从一个亲缘关系疏远的物种那里获得了 P 因子

(见第117页)

7.6 转座与物种形成

物种形成(speciation)或**分枝进化**(cladogensis)(即从一个亲本物种产生两个或多个物种)是最重要的进化过程之一。遗憾的是,在分子水平上,它也是我们了解得最少的进化过程之一。我们不知道新物种是通过什么方式而从老物种中产生的。我们所知道的只是,物种形成的过程需要在两个属于同一物种的群体间产生生殖障碍,使得它们就不再能杂交。杂种劣势有一阵子曾被认为是物种形成过程的早期阶段,起着属同一物种的不同群体间交配后产生生殖隔离的机制的作用。事实上,姐妹种 *D. melanogaster* 和 *D. simulans* 间杂交产生的杂种不育就与劣势非常类似(例如,性腺发育不良,分离扭曲等)。

不过,此观点有几个问题有待解决。第一,虽然杂种表现出适合度降低,因而从生殖角度看是部分

被隔离的，但 P 因子在种系细胞中的转座实际上保证了大多数传给杂种的染色体将带有 P 因子，且细胞型也最终会变成 P 型。于是，倘若杂种的适合度降低不是太大的话，则 P 因子将会传遍整个群体。事实上，因为有效的生殖隔离，杂种最终将是几乎完全不育的。第二，P 因子有作为传染因子而从一个个体向另一个个体水平转座的能力（见第117页原版）。于是，整个群体可能会被 P 迅速接管，这样杂种劣势看来在自然界中只能维持非常短的时期。事实上，许多果蝇种已知其所有个体和所有群体都带有 P 因子或类 P 因子，因而杂种劣势将不会在这些种中的任何一种中出现。最后，就我们所知，杂种劣势限制在果蝇范围，因而可能并不代表自然界中的普遍现象。即使在果蝇中，也没发现易动因子与阻碍姐妹种间基因流动有关的证据。

由于可转座因子的发现，许多其他关于由转座导至物种形成的机制也在文献中出现。例如，曾有人提出，一个群体中的含有调控序列的因子，它的大规模复制型转座可能会引起所谓基因组的**遗传重排**(genetic resetting)，藉此许多基因将会经受一种新的调控形式。这样一个群体显然将变得与保留着旧调控形式的群体发生生殖隔离。另一种说法则求助于机制不相容性 (mechanical incompatibility)，它也是由大规模复制型转座所造成的。在这种情况下，假定一个群体中的可转座因子曾发生增殖，并达到使染色体大小出现显著增大的程度。于是，从一个亲本那里继承大染色体而从另一个那里得到小染色体的杂种生物体，在减数分裂期间将会经历染色体配对上的困难，而这极有可能会造成不育。遗憾的是，迄今为止所提及的物种形成模型，没有一个曾得到经验数据的支持。

7.7 可转座因子拷贝数的进化动力学

一个基因组中的可转座因子的拷贝由3个因素所决定。(1)u，一个可转座因子产生一个新基因组拷贝的概率（即复制型转座的概率），(2)v，该因子被切离的概率，以及(3)对抗增加可转座因子在基因组中的数目的选择强度。关于果蝇 *D. melanogaster* 群体中几个可转座因子的 u 和 v 值，已经由实验得出，转座频率被发现在不同可转座因子间有些变化，但平均起来则处在每因子每世代10^{-4}这一数量级。切离频率大约要低一个数量级(Charlesworth 和 Langley，1989)。因此，在没有对抗可转座因子转座的选择的情况下，基因组中的拷贝数预期将无限制地增加。

如果可转座因子的数目维持在一个平衡位置上，这可能是一个在自然界不成立的假定，那么，选择就必须发生作用以对抗拷贝数的增加。在最简单的决定性模型中，我们假定，个体的适合度 w 将随拷贝数 n 的增加而降低。此假定的合理性是，可转座因子的插入频频地改变着邻近基因的表达；随着可转座因子的数目的增加，基因表达出现有害变化的概率也会增加。可以证明，只要 w 值随着 n 的增加而降低，则不管 n 和 w 间的精确关系如何，群体在平衡位置相对于缺少可转座因子的个体而言的平均适合度为

$$\bar{w} = e^{-n(u-v)} \tag{7.1}$$

(Charlesworth，1985)。

在果蝇 *D. melanogaster* 的情况中，有约50个转座子家族，每个家族平均在基因组中出现10次(Finnegan 和 Fawcett，1986)，于是 n=500。因为 v 比 u 至少小一个数量级，所以 $u-v \approx u = 10^{-4}$。解等式7.1，我们得 $\bar{w}=0.95$。于是，适合度上的降低为 $s=1-0.95=0.05$。由等式7.1给出的平衡的稳定性，要求适合度的对数衰减比 n 的线性增加更陡峭(Charlesworth，1985)。不过，为了运算简单，我们假定呈线性，那么，每增加一个转座子适合度的降低近似为 $0.05/500=10^{-4}$。这样小的选择系数本质上意味着，一个生物体中的可转座因子的拷贝数主要是由随机遗传漂变所决定的。如果一个生物体含有数目更大的可转座因子，甚至在果蝇中可转座因子数大大超过500的可能性也是存在的(Rubin，1983)，那么，一个转座子对适合度的效应也许要比上面所得到的更小。

关于对抗拷贝数增加的选择的一个可采用形式是自我调控转座机制，即转座的频率随拷贝数的增加而降低，或切离的频率随拷贝数的增加而增加(见 Chenlesworth 和 Langley，1989)。

7.8 水平基因转移

水平基因转移(horizontal gene transfer)被定义成遗传信息从一个基因组向另一个基因组,特别是在两个物种之间的转移。这个术语的提出是为了把这类转移与通常的"垂直转移"区别开来,后者则是亲代把遗传信息传给子代的转移。水平基因转移要求,(1)在各生物体和各细胞间转输遗传信息的交通工具,和(2)将外源 DNA 片段插入宿主基因组中的分子机制。反录病毒能完成这两个任务,因为它们既能把染色体的 DNA 掺和到它们的基因组中,又能跨越物种界线(Benveniste 和 Todaro,1976;Bishop,1981)。在转座子以及其他类型的 DNA 中介的转座中,跨越细胞的传输必须由一个传染因子来提供,如质粒。事实上,许多自然界中出现的质粒都含有可转座因子,这些因子可以从质粒上离开而向细菌染色体移动,也可相反方向地移动。

通过发现某一具体基因的系统发育分布中的显著不连续性,也许能检出一个水平基因转移事件。例如,细菌 *Salmonella typhimurium* 含有一个类组蛋白基因,据我们所知它在其他细菌中没有对应物(Higgins 和 Hillyard,1988)。当发现基因系统发育和物种系统发育之间明显地存在矛盾时,特别是顺序类似性看来反映了地理上的近似而不是系统发育上的亲缘性时,或许也可以怀疑有水平基因转移。作为例子,我们考虑一下图7－15a 中的系统发育树。假定 B 从 A 那里分岐后,发生了一个 DNA 片段从物种 B 向物种 C 的转移。在除发生了水平转移的基因外的任何基因的顺序比较的基础上,我们可望得到一个正确的物种间系统发育关系。对比之下,如果我们采用发生了水平转移的 DNA 片段,则我们将得到如图7－15b 中那样的错误树。不过,我们也注意到,除水平基因转移外还有别的因素也能造成物种树和基因树之间的矛盾(第五章)。

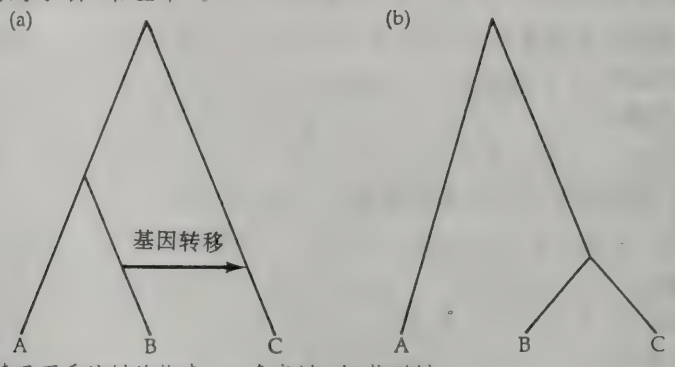


图7－15 水平基因转移的情况下系统树的构建。(a)真实树,(b)推测树。

两类序列能被水平地转移:(1)从可转座因子衍生出的序列,和(2)基因组序列。基因组序列的水平基因转移的例子中,很少有已被令人信服地证实了的。许多原来这样声称的例子,后来发现并不能得到分子证据的支持。而且,我们注意到,一个发生了水平转移的基因在宿主中保留其功能性的情况是很罕见的,预期比一个在同一物种内从一个基因组位置到另一个位置转移的基因(见第111页)更少出现。

病毒基因从狒狒到猫的水平转移

脊椎动物基因组中含有许多与反录病毒同源的序列。这些序列是真核生物细胞核 DNA 的正常组份,称为**内源性反录病毒序列**(endogeneous retroviral sequences)或**病毒基因**(virogene)。有几个内源性反录病毒序列在脊椎动物的种间转移的例子(综述见 Benveniste,1985)。其中一例与来自狒狒的 c 型病毒基因有关(图7－16)。

与狒狒的病毒基因同源的序列已在所有古世界猴的细胞 DNA 中检出。它们间的顺序类似性是与物种间的分类学关系紧密相关的。于是,该病毒基因在灵长类中至少存在了3千万年。有趣的是6种与家猫(*Felix catus*)亲缘关系密切的猫也含有这一序列,虽然在亲缘关系较远的猫科动物、象狮、豹和短尾猫,以及任何其他食肉动物中,都不存在这一序列。所以,这一序列极有可能是在过去某一时刻在物种间水平转移的。水平传递的年代和方向可从两类资料中推出:(1)顺序类似性和(2)古地质学信

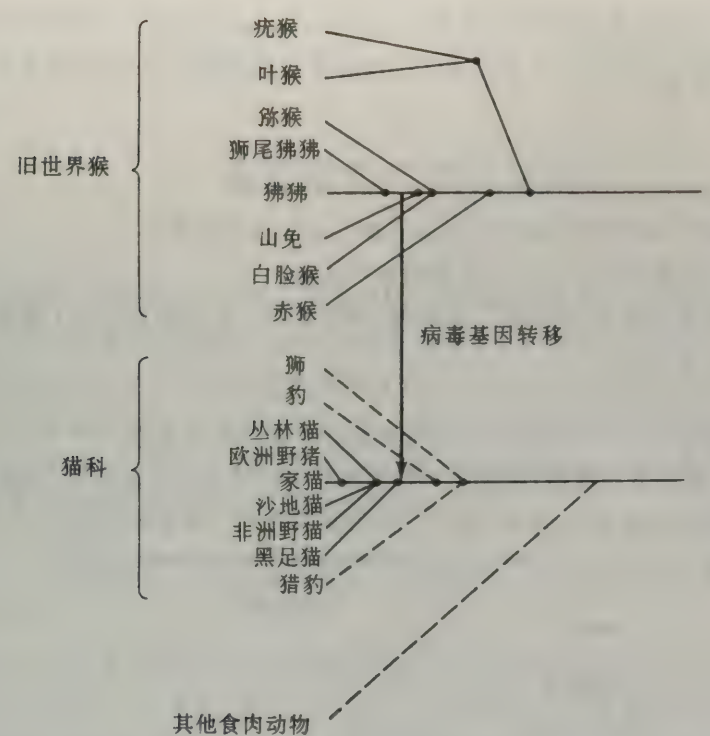


图7－16　古世界猴和猫的系统发育树。含有c型病毒基因的物种用实线标出。来自狒狒和狮尾狒狒的c型病毒基因展现出与猫的c型病毒基因有最大的类似性。因此，在大约1千万前可能发生了一次水平基因转移，该转移是从狒狒和狮尾狒狒的祖先到与狮、豹和猎豹等的谱系分歧后的现代猫的祖先。自 Benveniste (1985)修改而成。

息。

所有含狒狒病毒基因的猫种都来自地中海地区，而来自东南亚、新世界和非洲的猫科动物则缺少该序列。因此，转移发生在猫科动物的大辐射之后，且被限制在一个地理区域中。这一结论指出，水平基因转移的年代在500－1000万年前。传递方向则可这样推出：一方面考虑该序列在灵长类中的分布，另一方面考虑它在猫类中的分布。因为所有古世界猴都具有该病毒基因，而只有几种猫才具有该基因。所以，只能假定猫类是从狒狒那里获得该序列，而不是相反，这才是合理的。这一结论在考虑到以下事实后得到强化：猫类与3种狒狒(*Papio cynocephalus*，*P. papio* 和 *P. hamadryas*)和亲缘关系密切的狮尾狒狒(*Theropithecus gelada*)中的病毒基因，它们的相似程度比猫类与任何其他灵长类序列的相似程度要更高一些。因此，该序列一定是狒狒和狮尾狒狒与山魈分岐后不久而从它们的祖先转移到猫类中去的(图7－16)。从对狒狒的研究中得出的年代与从对猫类的研究中得出的年代一致得相当好。

P 因子在果蝇的种间的水平转移

水平基因转移的另一个例子与 *D. melanogaster* 中的 *P* 因子有关。正如前面所讲的，在近40年的时间里 *P* 因子已迅速传遍了 *D. melanogaster* 的自然群体(见第114页)。而 *P* 因子不存在于与 *melanogaster* 亲缘关系很近的一些种，象 *D. mauritania*，*D. sechellia*，*D. simulans* 和 *D. yakaba* 等之中。那么这些因子是从哪里来的呢？丹尼尔斯等(Daniels 等，1990)曾对成百种果蝇进行了普查，结果表明，除 *D. melanogaster* 外，*melanogaster* 类亚组的任何其他种中都未发现 *P* 序列。相反，亲缘关系疏远的 *willistoni* 组和 *saltans* 组的所有种都含有 *P* 因子和类 *P* 因子。尤其是，*D. willistoni* 的 *P* 因子被发现除了一个碱基替换外完全等同于 *D. melanogaster* 中的因子，这指示 *D. willistoni* 在 *P* 因子转到 *D. melanogaster* 的水平基因转移中，起着供体物种的作用。

有几个原因使得我们怀疑这种水平基因转移是最近才发生的。首先，来自 *D. melanogaster* 和来自 *D. willistoni* 的 *P* 序列间的接近等同，表明发生分岐的时间非常短。其次，来自地理位置相距很远处的 *D. melanogaster*，它们的 *P* 序列间几乎没有遗传变异性，这指示自 *P* 因子导入 *D. melanogaster* 以来时

间还很短，以至还来不及积累遗传变异性。最后，*P* 因子在 *D. melanogaster* 中出现的地理模式，以美洲大陆中的群体最先获得它，看来它指示着涉及了一次非常近期的侵入事件，时间或许在近50年之内。

习题

1. 反录因子和反录序列间的差异是什么？

2. 给出一个基因组序列，你怎么能讲出它是不是一个反录序列呢？

3. 怎样才能将加工后假基因和未加工假基因加以区别？

4. 大多数加工后假基因都是"垂死者"。请解释，为什么这一条件使它们成了推测点突变模式的极好材料（见第四章和 Li 等，1984）。

5. 解释为什么反录基因很少见。

6. 为什么 *Alu* 序列在人类和其他灵长类的基因组中含量会如此丰富？

7. *Alu* 序列曾被认为是有功能的，但现在却普遍相信它们是加工后假基因。在此假说下，*Alu* 序列应该象其他假基因那样迅速地进化。那么，表7－3中的数据与这一假说相容吗？

表7－3　*Alu* 序列之间以及 η－珠蛋白假基因之间顺序岐化的程度

物种对	百分比岐化度	
	Alu 序列[a]	η假基因
人对黑猩猩	2.2±1.4	1.7
人对马来猩猩	3.7±1.9	3.1

数据自 Koop 等（1986a）和 Li 等（1987a）

a. 7个垂直相关的序列被用来计算平均值和标准偏差。

8. 试列举转座对宿主可能产生的有利和不利效应。

后继阅读文献

Berg. D. E. and M. M. Howe (eds). 1989. *Mobile DNA*. Academic Society for Microbiology, Washington DC.

Compbell. A. 1983. Transposons and their evolutionary significance. pp. 258－279. In M. Nei and R. K. Koehn (eds). *Evolution of Genes and Proteins*. Sinauer Associates, Sunderland. MA.

Charlesworth. B. and C. H. Langley. 1989. The population genetics of *Drosophila* transposable elements. Annu. Rev. Genet. 23: 251－287.

Doolittle. R. F., D-F. Feng, M. S. Johnson and M. A. McClure. 1989. Origins and evolutionary relationships of retroviruses. Quart. Rev. Biol. 64: 1－30.

Lewin. B. 1990. *Genes IV*. Oxford University press. Oxford.

Scaife. J., D. Leach and A. Galizzi (eds.). 1985. *Genetics of Bacteria*. Academic Press, New York.

Shapiro. J. A. (eds). 1983. *Mobile Genetic Elements*. Academic Press. New York.

Varmus. H. 1988. Retroviruses. Science 240: 1427－1435.

Weiner. A. M., P. L. Deininger and A. Efstratiadis. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. Annu. Rev. Biochem. 55: 631－661

8 基因组的组织化和进化

关于基因组的组织化和进化的讨论应该包括3个不同的课题。第一个是基因组大小，这在不同生物间变化极大。这种变化是如何形成的？增加和降低基因组大小以产生这种变化的机制是什么？第二个是基因组中的遗传信息。基因组主要是由基因的 DNA 构成的吗？还是主要是由非基因的序列构成的？基因组中有许多重复序列吗？而如果是这样那么重复序列的染色体分布模式是什么？非基因部分有功能吗？或者它只是“废物”？第三个课题涉及基因组的核苷酸组成。基因组的不同区域中在组成上有异质性吗？什么机制使核苷酸在组成方面造成区域性差异？

8.1 C 值

单倍体基因组，象在精子细胞核中的那样，其 DNA 的量称基因组大小(genome size)或 C 值(C value)，这里 C 来自“constant”(常数)或“characteristic”(特征)，表示单倍体基因组的大小在任何一个物种中都是相当恒定的这样一个事实。反之，C 值在原核生物和真核生物的各物种间则变化幅度极大。

表8－1 常用于测试基因组大小的单位换算

单位	换算因子		
	微微克	道尔顿	碱基对
微微克	1	6.02×10^{11}	0.98×10^{9}
道尔顿	1.66×10^{-12}	1	1.62×10^{-3}
碱基对	1.02×10^{-9}	618	1

真核生物中细胞核基因组的大小通常用微微克(pg)DNA 来度量($1pg=10^{-12}g$)。较小一些的原核生物的基因组则通常用道尔顿来度量，道尔顿即一个原子或分子的相对质量单位。还有更小的基因组，象细胞器和病毒等的基因组，以及某些特殊的 DNA 延伸物，它们的大小则常用双链 DNA 或 RNA 的碱基对(bp)或千碱基对(kb)来表示(1kb＝1000bp)。为避免混乱，我们将只用 bp 和 kb。单位换算因子可见表8－1。

8.2 细菌的基因组大小的进化

细菌的基因组大小其变化幅度超过了20倍，从某些限制性细胞间寄生物中的约6×10^{5}bp 到几种蓝细菌中的10^{7}bp 以上(表8－2)。最小的游离生活的原核生物，枝原体，含有包括50种核糖体蛋白在内的350个为蛋白质编码的基因，两组 rRNA 基因(5S，16S 和23S)，和大约40个 tRNA 基因。因此，一个枝原体的基因组由大约400个基因所构成，这接近于所猜测足够维持机体自主性生活的最小数目(Muto 等，1986)。其他细菌中的基因数大致在500到8000的范围内变化，换句话说，在特征细菌的种类中，基因数上的变异与 C 值方面的变异相同。因此，细菌看来不会含有大量的非基因的 DNA。

细菌的基因组可被划分成3个部分：(1)染色体 DNA，(2)源于质粒的 DNA，和(3)可转座因子(Hartl 等，1986)。染色体部分含有为生长和代谢功能所必需的蛋白质编码基因(70－80%)、间隔序列和各种信号序列(20－30%)、编码 RNA 的基因(～1%)，以及一些通常长度在几十个碱基对这一级别上的短重复序列。细菌常带有许多作为染色体外遗传成份的质粒，然而，在有些例子中，一些来自质粒的基因却发现被整合在细菌染色体中。可转座因子是某些细菌基因组的普通组分。例如，*E. coli* 的

染色体，至少含有6个不同类型的插入序列各1－10个拷贝（第七章）。基因组的非染色体部分（包括染色体中的插入序列和来自质粒的基因）看来要比染色体部分小一个数量级。

表8－2　细菌中C值的范围

分类单位	基因组大小的范围(kb)	比例(最高/最低)
真细菌	650－13200	20
革兰氏阴性	650－7800	12
革兰氏阳性	1600－11600	7
蓝细菌	3100－13200	4
枝原体	650－1800	3
古细菌	1600－4100	3

资料自 Cavalier-Smith(1985)。

细菌中基因组大小的分布是不连续的，在模值约0.8×10^6，1.6×10^6和4.0×10^6bp处呈现出主峰，而在7.2×10^6和8.0×10^6bp处有几个副峰(Herdman，1985)。这种分布引出了这样的看法：较大的基因组是从较小的基因组经连续循环的基因组重复进化而来的。由于基因组大小和细菌的系统发育间看来没有显著关系，所以，已经出现了这样一个提议：基因组重复在细菌谱系的进化中是频频出现的(Wallace 和 Morowitz，1973)。

应用根据 rRNA 顺序比较而得出的尝试性细菌系统发育，赫德曼(Herdman，1985) 曾找出了基因组大小变化与系统发育史的关系。其结果指示，基因组重复在不同细菌谱系中是独立地发生的。有趣的是，许多重复看来曾在进化中的某一相当特别的时刻，即氧在大气层中出现后不久，大约在18亿年前，同时出现在几种不同的细菌谱系中的。

概括起来，细菌中基因组大小的分布能用几个过程的结合来加以解释：(1)某些谱系中呼吸代谢独立进化期间的基因组重复，(2)独立地出现在许多谱系中的，随后的基因组重复。(3)小规模的缺失和插入，(4)复制性转座，(5)主要来自质粒的基因的水平转移，和(6)许多寄生性谱系中大量 DNA 丢失。

8.3　真核生物的基因组大小和C－值悖论

真核生物中的C值通常比原核生物中的要大得多，但也有例外。例如，酵母 *Saccharomyces cerevisiae* 的基因组就比许多革兰氏阳性菌和大多数蓝细菌的小。因为真核生物的基因组有多重的复制原点，而大多数原核生物看来只有一个，所以，在复制出较小原核生物基因组所需的相同的时间里，真核生物复制出的 DNA 的量就要大得多。

真核生物中的C值变化要比细菌中的大得多，从8.8×10^6bp到9.9×10^{11}bp，范围接近80000倍（表8－3）。单细胞的原生生物，特别是肉足类变形虫，在C值上表现出的变化最大。3个羊膜类的纲（哺乳类、鸟类和爬行类）则相反，在真核生物中以它们在基因组大小上的变化小（最高相差4倍）而显得不同寻常。其他的纲，对那些C值数据大体上存在者而言，表现出至少100倍的变异。

有趣的是，真核生物间基因组大小上的巨大种间差异看来与生物的复杂性和生物所编码的可能基因数没有什么关系。例如，几种单细胞的原生动物具有比哺乳动物要多得多的 DNA（表8－4），而后者据推测应要更复杂一些。而且，一些复杂性看来相似的生物（例如，果蝇与蝗虫，洋葱与百合，双核草履虫（*Paramecium aurelia*）与尾草履虫（*P. caudatum*）却展示出极大的C值差异（表8－4）。这种C值与基因组中所含的遗传信息的认定数量间缺乏对应的现象，以**C值悖论**或**C值矛盾**(C value paradox)而著称于文献之中。在姐妹种（即一些相互在形态学上极为相似，以至表型上不可区分的种）的比较中C值矛盾也是明显的。在原生生物、真骨鱼类，两栖类和显花植物中，许多姐妹种，尽管根据姐妹种的定义这些生物间的复杂性没有什么差异，但在它们的C值上却差别极大。我们可以假定一个物种不可以具有少于保证其生命功能所必需的 DNA 的量，但我们必须解释为什么有那么多物种含有远远过剩的 DNA 的量。

表8—3　不同真核生物类群中的C值范围

分类单位	基因组大小范围(kb)	比例(最高/最低)
原生生物	23,500—686,000,000	29,191
眼虫类	98,000—2,350,000	24
纤毛(亚门)类	23,500—8,620,000	367
肉足(总纲)类	35,300—686,000,000	19,433
真菌	8,800—1,470,000	167
动物	49,000—139,000,000	2,837
海绵	49,000—53,900	1
环节动物	882,000—5,190,000	6
软体动物	421,000—5,290,000	13
甲壳类	686,000—22,100,000	32
昆虫	98,000—7,350,000	75
棘皮动物	529,000—3,230,000	6
无颌类	637,000—2,790,000	4
鲨与苔	1,470,000—15,800,000	11
真骨鱼类	382,000—139,000,000	364
两栖类	931,000—84,300,000	91
爬行类	1,230,00—5,340,000	4
鸟类	1,670,000—2,250,000	1
哺乳类	1,420,000—5,680,000	4
植物	50,000—307,000,000	6,140
藻类	80,000—30,000,000	375
蕨类植物	98,000—307,000,000	3,133
裸子植物	4,120,000—76,900,000	17
被子植物	50,000—125,000,000	2,500

资料自 Cavalier-Smith(1985)及其他来源。

要搞清楚的第一个问题是，基因组大小和基因的数目间是否存在相关。换言之，基因组大小上的差异应归因于基因的DNA还是非基因的DNA?真核生物在为蛋白质编码的基因数上表现出约有50倍的变异，从酵母中的约3000个到哺乳动物中的150000个左右(Cavalier-Smith，1985)。这种50倍的差异显然不足以解释细胞核DNA在含量上的80000倍的变异。再者，基因数是与结构复杂性正相关的，而基因组大小则不是。mRNA分子长度上的种间差异也不能解释C—值矛盾。在不同生物间编码区和非编码区的平均长度都有着差异的情况下，基因长度和基因组的大小间就不会存在相关。

几种编码RNA的基因的重复程度与基因组大小间的正相关业已被发现(第六章)。类似地，象端粒、着丝粒和复制器基因，这样一些在有丝分裂和减数分裂期间为染色体复制、分离和重组所必需的调控序列，它们的拷贝数与基因组大小间也存在相关性。不过，所有这些基因只构成基因组的一小部分，以至RNA基因和调控序列等的在数量上的差异并不能解释基因组大小上的差异。

总之，我们只剩下非基因DNA部分作为造成C—值矛盾的唯一嫌疑犯。换句话说，真核基因组的主要部分是由不含遗传信息的DNA所构成的。曾有估计认为，每基因组非基因DNA的量在真核生物中从约3.0×10^{6}bp到1.0×10^{11}bp以上(10万倍范围)之间变化，且构成了基因组的量的不到30%至几乎100%(Cavalier-Smith，1985)。

表8－4　按基因组大小等级的真核生物的C值

物种	C值(kb)
Navicola pelliculosa(硅藻)	35,000
Drosophila melanogaster(果蝇)	180,000
Parameciunm aurelia(纤毛虫)	190,000
Gallus domesticus(鸡)	1,200,000
Erysiphe cichoracearum(真菌)	1,500,000
Cyprinus carpio(鲤)	1,700,000
Lampreta planeri(七鳃鳗)	1,900,000
Boa constrictor(蛇)	2,100,000
Parascaris equorum(蛔蚓)	2,500,000
Carcarias obscurus(鲨)	2,700,000
Rattus norvegicus(大鼠)	2,900,000
Xenopus laevis(蛙)	3,100,000
Homo sapiens(人)	3,400,000
Nicotiana tabaccum(烟草)	3,800,000
Paramecium caudatum(纤毛虫)	8,600,000
Schistocerca gregaria(蝗虫)	9,300,000
Allium cepa(洋葱)	18,000,000
Coscinodiscus asteromphalus(硅藻)	25,000,000
Lilium formosanum(百合)	36,000,000
Amphiuma means(蝾螈)	84,000,000
Pinus resinosa(松)	68,000,000
Protopterus aethiopicus(肺鱼)	140,000,000
Ophioglossum petiolatum(蕨类)	160,000,000
Amoeba proteus(变形虫)	290,000,000
Amoeba dubia(变形虫)	670,000,000

资料自 Cavalier-Smith(1985),Sparrow 等(1972)和其他参考文献。

8.4　真核生物基因组的重复结构

真核生物的基因组有两大特征:(1)序列的重复,和(2)组成分隔化,分隔成以特别的核苷酸组成而相互区别的特征片段。重复 DNA(repetitive DNA) 是由不同长度和组成的核苷酸序列所构成,这些序列以串接或散在的形式在基因组中几次出现。不发生重复的 DNA 片段被称为**单拷贝 DNA**(single-copy DNA)或**单一 DNA**(unique DNA)。基因组中由重复序列占据的部分在各分类单位间变化极大。在酵母中,该部分的量约为20%;而在哺乳类中,则高达60%的 DNA 是重复性的。在植物中,该部分可超过80%,而且比它更高的值也曾被记录到(Flavell,1986)。

布里滕和科恩(Britten 和 Kohne ,1968) 所做的关于 DNA 双链重新结合的经典研究表明,较高等的真核生物,其基因组大致分分成4部分:**折回 DNA**(foldback DNA),**高度重复 DNA**(highly repetitive DNA),**中度重复 DNA**(middle-repetitive DNA) 和**单拷贝 DNA**(图8－1)。折回 DNA 由回文 DNA 顺序所构成,一旦这种变性了的 DNA 被允许复性时,该顺序可以形成发夹式的双链结构。高度重复部分是由长度从几个到成百个核苷酸的短序列所构成,其重复数平均为50万次。中度重复部分由平均达成百或成千碱基对的长序列所构成,它们在基因组中以成百的次数出现。

根据重复在基因组中散在的模式,重复部分已被发现是由两类重复家族所构成的:区域性重复家

族和散在的重复家族。

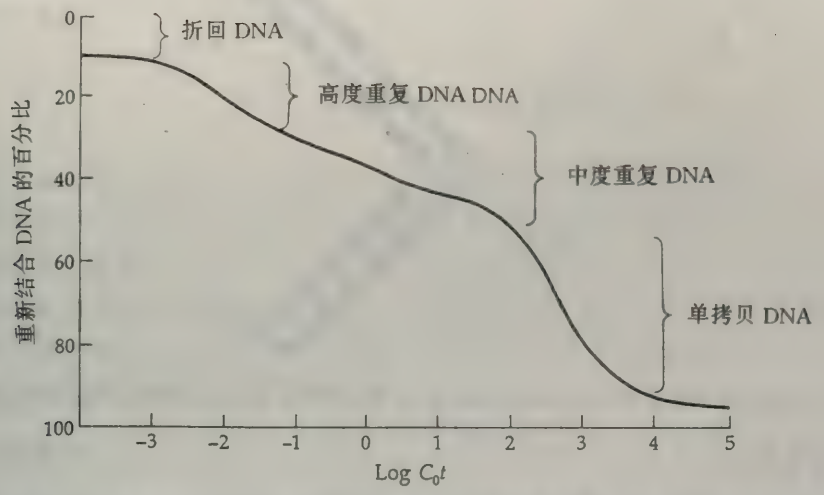


图8－1　哺乳类的 DNA 重新结合图。DNA 被纯化、裁剪，加热熔解成单链，然后通过逐渐冷却而使之重新结合。纵轴上的重新结合的双链 DNA 百分比，显示出是 DNA 的浓度与横轴上的时间(Cot) 之积的函数。自 Schmid 和 Deiniger (1975)修改而成。

区域化的重复序列

大多数真核生物的基因组含有串联排列的、高度重复的 DNA 序列。在有些物种中，这些区域化的重复 DNA 序列可构成基因组中 DNA 的主要部分。例如，在袋鼠 *Dipodomys ordii* 中，基因组的50%以上由3个重复序列构成：AAG(24亿次)TTAGGG(22亿次)和 ACACAGCGGG(12亿次)见(Widegren 等，1985)。当然，这些家族并不是完全同质的，而是含有许多变异型，它们与一致的顺序有一个或两个核苷酸差异。例如，TTAGGG 家庭中，有些序列实际上是 TTAGAG。虽然如此，但许多区域化的高度重复序列仍有着相当一致的核苷酸组成，以至根据基因组 DNA 的部分化和密度梯度分离，它们能形成一条或多条粗带。这些带明显地与 DNA 的主带和由其他组成更异质的片段产生的带有区别，被称为**卫星 DNA**(satellite DNA)。

在有些物种中，所有染色体上都发现有串接排列的高度重复序列，而在另一些物种中它们则被限制在某一特别的染色体位置中。例如，果蝇 *D. nasutoides* 基因组的60%由卫星 DNA 构成，而这类 DNA 全部都处在4对染色体中的一对上(图8－2)，而这种染色体上看来几乎不含别的 DNA(Miklos，1985)。

根据当前可得到的证据，区域化的高度重复序列很有可能是无功能的。而且，区域化的高度重复序列可能既不会降低、也不会提高个体的适合度。因此，这类序列的进化不受自然选择的影响，而主要是由基因转变和不等价交换(第六章)所决定的。这些机制将产生两种结果：(1)序列同质性，和(2)拷贝数随进化时间而大幅度波动(见 Charlesworth 等，1986)。还曾经有这样的提法：区域化的重复序列的翻转速率是相当高的，即，现存的排列可能会通过不等价交换而被排除，而新的排列则可能会通过 DNA 重复过程而不断地产生(Walsh，1987)。

大多数串接重复序列只不过是"废物"DNA 的说法，本质上是说，它们没有表型效应或者它们对它们的携带者的适合度没有影响。这在大多数情况下可能是真实的，但也有证据表明高度重复序列的某种特殊排列并不总是如此。果蝇 *D. melanogaster* 的自然群体中，应答者 *Responder*(*Rsp*)基因座位由20－2500个拷贝的富含 AT、长120bp 的序列所构成(Wu 等，1988)。有一次竞争性实验涉及某一混合群体，该群体由含700个重复拷贝的果蝇和只含20个拷贝的果蝇构成，实验中观察到只含20个重复的果蝇频率随时间推移而降低(Wu 等，1989)。因此，已有结论是，含700个拷贝的果蝇有比只含20个拷贝的果蝇更高的适合度。目前，*Rsp* 的功能尚不知，但它显然不是"废物"DNA，因为它的缺乏会降低适合度。这些结果能否用于其他串联重复的序列，特别是用于主要卫星 DNA 家族，也是未知的。

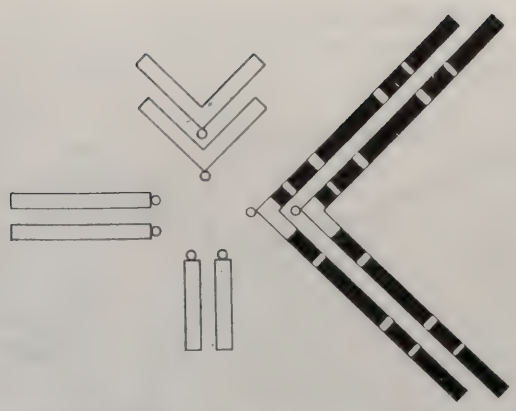


图8－2　果蝇 *D. nasutoides* 中高度重复 DNA 序列(黑色区)位于一条染色体上。自 Miklos(1985)修改而成。

散在的重复序列

第二类高度重复的 DNA 由散在于整个基因组中的序列所构成。散在的高度重复序列的拷贝已在内含子、基因的侧翼区域、基因间区域、和非基因 DNA 中发现。有两大类散在的高度重复序列：短的散在重复序列，缩写成 SINE，和长的散在重复序列，或 LINE(Singer，1982)。

SINE 中的典型代表为短于500bp，且在单倍体基因组中出现拷贝数在10^5或10^5以上者。与确定 tRNA 的基因一样(第一章)，许多 SINE 含有内部转录信号且被 RNA 多聚酶Ⅲ转录。大多数 SINE 是反录序列。人的基因组中最著名的 SINE 家族是 *Alu* 家族(第七章)。

LINE 原来曾被描绘成长于5kb、且每基因组以10^4或其以上的拷贝数存在的 DNA 序列(Singer，1982)。人的基因组中显然只含有一个 LINE 家族、即 L1(Hutchison 等，1989)。这种一致性的 L1序列长约6kb，在一端有一个 poly－A 尾巴，且一般由长度短于20bp 的短顺向重复构成侧翼。拷贝数大约为每单倍体基因组10万个。大多数 L1序列在它们的5′端处被截断。完整的 L1序列含有2个长开放阅读框架，ORF－1和 ORF－2，分别有约375和1300个密码子。ORF－2含有的氨基酸序列具有反转录酶的主要特征。以上特点表明，L1因子是通过多聚腺苷化的 mRNA 的反转录，和该反转录本随后插入基因组中而产生的。由于 L1序列不具长末端重复(LTR)，所以它们很可能是反转录子(第七章)。大多数 L1序列不被转录。不过其中少数则用 RNA 多聚酶Ⅱ进行转录。

与人的 L1家族同源的 LINE 已在所有哺乳类，包括有袋哺乳类中发现(见 Hutchison 等，1989)。与 L1相关的非哺乳类因子包括，果蝇中的 I、F、G 和 D 因子，*Trypanosoma brucei* 中的 *Ingi*，家蚕 *Bombyx mori* 中的 R 2，和玉米中的 *Cin* 4。物种间的 L1序列分岐程度远大于同物种的 L1拷贝间的分岐程度。例如，来自小鼠和人类的 L1序列相互平均约有30％的差异。相比之下，小鼠中 L1因子间的顺序分岐量只有4％(Hutchison 等，1989)。

由于大多数 L1序列被截去尾部，所以它们不含有完整的阅读框架。结果，它们也许不能转座。这类有缺陷的因子被发现比完整的因子进化得更为迅速。而且，有缺陷 L1序列的进化谱系已被发现不含分枝，这指示这些因子不能复制性转座。于是，它们可被看成是反转录子的假基因，对它们而言功能限制已不再有作用。大多数 L1序列是有缺陷的，这一事实的含义是，基因组中的 L1因子的传播仅依赖于少数的源因子。结果，基因组中的 L1因子是高度同质的，且序列翻转的速率非常高。事实上，曾有估计，啮齿类中一半以上的 L1因子只有300万年或者更短一些的历史(Hardies 等，1986)。

哈奇森等(Hutchison 等，1989)曾建议，SINE 和 LINE 应该重新加以定义。即，不用长度和拷贝数作诊断特征，而把 LINE 看成是，应包括所有具有为中介它们自己和它们后代的转座的蛋白质编码活性的反转录子和反录转座子。另一方面，SINE 则应包括不为这类功能编码的反录序列。

可转录基因的基因组位置

RNA－DNA 杂交实验表明，只有很小一部分的 可转录基因位于基因组的重复部分中。大多数可转录基因处在单－DNA 部分内。即使在单－部分中，大多数序列也是不转录的。事实上，人类中只有

3%的非重复 DNA 序列被转录(见 Lewin,1990)。这些实验进一步地支持真核生物基因组中大多数是不具遗传信息的观点。

8.5 增加基因组大小的机制

为了解释真核生物的基因组中非基因 DNA 大量存在的现象,我们首先必须探讨可能会导致基因组大小增加的过程。设想出来的机制应该不仅能解释非基因 DNA 本身的现象,而且能解释它的重复性和特别的基因组分布。

我们把基因组增加分成两种类型:(1)全局性增加,即整个基因组或它的主要部分,如染色体,重复,和(2)局部性增加,即某一具体序列增殖而产生重复 DNA。在后一种情况中,我们又分散在的重复序列产生的机制和区域化重复序列产生的机制。

基因组重复

由于真核生物的基因组显著地大于真细菌的基因组,所以,真核生物从原核生物的祖先进化而来的过程必须涉及基因组大小上的增加。有几种分子机制可以导致基因组大小上的增加。造成基因组增长的一个重要因素是基因组重复(genome duplication)或基因组加倍(genome doubling)。基因组重复是在 DNA 复制之后,由于缺少所有染色单体间的分裂过程而产生的结果。

假定哺乳类基因组比细菌的基因组大1000倍左右,并假定基因组重复是造成基因组增大的唯一原因,则我们可以推出,基因组从原始的细菌大小增加到现在哺乳类中的大小,大约需要10轮基因组重复。从另一个角度看,即基因组重复平均每3亿年出现一次。另一方面,如果 DNA 含量是以小段 DNA 连续增加的方式增长,比如说通过转座或不等价交换,则从细菌到哺乳类基因组增长的速率近似为每年7个核苷酸(Nei,1969)。

基因组大小的多峰态值的分布也已在许多真核生物类群中被记录到(Rees 和 Jones,1972;Grime 和 Mowforth,1982),这与细菌中的情形类似(见第119-120页)。这在单子叶植物中特别明显,它们的基因组大小在0.60,1.18,2.16,4.51和8.53×10^9bp 处出现峰值,表现出一种多峰态值的分布(Sparrow 和 Nauman,1976)。类似的分布也曾在别的类群,如棘皮动物、昆虫和真菌中观察到,以及在两栖类和真骨鱼类中,较小范围地观察到。基因组重复或多倍性重复(polyploidy)看来是真核生物基因组大小进化的主要机制。有趣的是,每一轮基因组重复都牵涉 DNA 的少量丢失,所以,每轮重复后 DNA 的量都是以一个略小于2的因子增加(Sparrow 和 Nauman,1976)。在新近形成的多倍体中,还谈不上 C 值方面的增加,因为该值是指单倍体的大小,而这两个基因组经历过突变、易位、染色体重排和染色体数目上的变化后,它们将最终会变成一个新的基因组,到那时才谈得上 C 值增加的问题。换言之,一个古老的多倍体与一个二倍体将是不可区别的(Cavalier-Smith,1985)。

多倍性重复是自然界中的一个普遍现象;然而,在进化的历史中多倍体看来很少能生存下来。在许多情况下,其原因是,多倍性重复是有害的,因而会受到选择的强烈对抗。多倍性重复的有害效应包括:(1)细胞分裂时间大大延长,(2)细胞核的体积增大,(3)在减数分裂的分离期间染单体断裂的数目显著增加,(4)遗传不平衡,和(5)在生物的性别由性染色体和常染色体间的数目比(例如果蝇),或性染色体与倍数性间的数目比(例如膜翅目)来决定的情况下,干扰性别分化。在有些情况下,多倍性重复可能是有利的,象在显花植物中,它可以减少或消除杂种不育性(见 Cavalier-Smith,1985)。

染色体重复

一个染色体的重复,或非整数倍重复,大多数是有害的。在哺乳类中,它常伴随着致死性或不育性。在人类中,众所周知的例子包括象唐氏(Down)综合症(21三体)和第18染色体三体等频频出现的病症。类似的有害症状也会与染色体的部分重复相关联。因此,染色体重复预期对基因组大小上的增加不会有显著贡献。

在许多生物中,特别是在那些被调查得最彻底的植物群体中,已经观察到几个建立了额外染色

体，即所谓 B-**染色体**(B-chromosomes)的例子(Jones，1985)。B-染色体不是整个染色体精确地重复，而可被看成是部分染色体重复的产物。这些额外染色体看来不含太多基因，而且它们大多数是由短重复序列所构成的(即它们是高度异染色质的)。因此，部分非整数倍重复可能在进化期间的基因组大小增长过程中起作用，但对于基因数的进化或许并无重大贡献。

基因组大小的区域性增加

基因组大小的区域性增加可由转座(第七章)和不等价交换(第一和第六章)所引起。前一种模式将产生散在的重复序列；后一种则产生串联的重复序列。基因组大小的区域性增加也可通过获得外来 DNA(第七章)而实现；不过，此过程的贡献与整个 DNA 大小比起来或许可忽略不计。

曾有建议认为，真核生物中所有中度重复的 DNA 部分都起源于可转座因子。不过，这些因子中大多数已不再可动，它们的转座能力已被突变或别的因子的插入摧毁了。例如，果蝇中的异染色质的大多数，实际上也许是这类已经死亡了的因子的墓地。不过，非活性的易动因子能通过不等价交换那样的过程而局部地增殖。

许多区域性重复序列曾经是由不等价交换而产生的。然而，看起来不等价交换更常做的是打散串联排列、而不是增加它们的大小和拷贝数，所以，该过程并不能解释所有区域性重复 DNA 存在的原因(Walsh，1987)。再者，不等价交换通常产生的是由相对而言较长的重复单位构成的序列。相比之下，许多区域性重复序列，象卫星 DNA，则是由非常短的，简单地重复的主体所构成的(见第123页)。为了解释这类序列存在的原因，DNA 扩增被提了出来。**DNA 扩增**(DNA amplification)是指，任何一种将某一基因或 DNA 序列的拷贝数增加到远超出一个生物的特征水平之上的事件。狭义地，DNA 扩增指在一个生物的寿命期间中发生，且使某一 DNA 序列的拷贝数突然增加的事件。

扩增的最有效方法之一是 DNA 复制的**滚环模型**(rolling-circle mode)(图8－3；Bostock，1986)。这

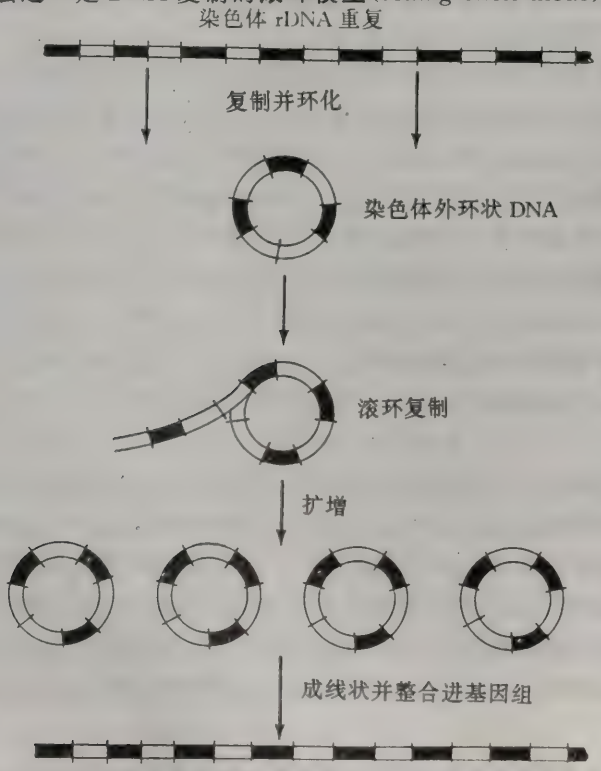


图8－3　两栖类卵母细胞中基因扩增的滚环模型。染色体的 rRNA 基因呈串联排列，含有可转录部分(黑色)和不转录部分(白色)。扩增涉及一个染色体外环状拷贝的形成，该拷贝含有数目可变的重复，通过多轮的滚环复制而扩增。注意，重复的周期性可能随滚环扩增而变化。自 Bostock(1986)修改而成。

种复制类型在两栖类卵母细胞中被用来进行 rRNA 基因的扩增。在这种情况下的扩增包括一个 DNA 序列的染色体外环状拷贝的形成，然后可产生许多新增加的含原串接重复序列的染色体外单位。如果这样一些单位整合回染色体中，则该基因组将增加了由等同的重复序列构成的部分。

复制滑脱(replication slippage)或**滑脱链误配**(slipped-strand mispairing)是这样一种过程,即DNA多聚酶转回来并用同一个模板以产生一个重复(图1—8)。现存的串接重复序列特别容易造成复制滑脱,因而此过程能产生短重复的非常长串的排列。滚环复制和复制滑脱都能提供关于串接重复序列在基因组中迅速增生的机制。然而,对于这些过程经验性证据仍然是有限的。

8.6 非基因DNA的维持

解决C—值矛盾问题以及说明真核生物基因组的结构,需要我们为大量非基因的、看起来完全是多余的DNA的长期维持提出一个进化机制。这反过来又与这样的问题紧密联系在一起:这种DNA如果有功能的话、那么可能会是什么样的功能呢?为对这一现象作出进化论的解释,已有不少尝试。下面我们将给出4种这样的假说。

1、非基因DNA行使着至关重要的功能,象基因表达和全局性调控之类(Zuckerkandl,1976)。根据这一假说,DNA的过剩只是表面上的,而这种DNA是全部都有功能的。因此,这类DNA的缺失或去除将对适合度具有害影响。

2、非基因DNA是无用的废物DNA(Ohno,1972),它被染色体所被动地携带仅仅只是因为它与有功能的基因在实体上是连续的。根据这一观点,则过剩的DNA并不影响该生物的适合度,于是将无限制地从一代传给另一代。

3、非基因DNA是一种无功能的"寄生物"(Ostergren,1945)或"自私DNA"(Orgel和Crick 1980;Doolittle和Sapienza,1980),是通过基因组间选择而累积且被积极地维持的。

4、DNA有一种结构的或核类型的(nucleotypic)功能,即一种与携带遗传信息的任务无关的功能(Cavalier-Smith,1978)。

关于第一种假说的证据极少。事实上,大多数资料倒是指示出,现在被认为是非基因的DNA事实上就是无功能的,它们中大多数可以缺失而不造成可查觉的表型效应。真核生物中的过剩DNA看来也不会加重代谢系统的负担,在维持和复制大量非基因DNA时能量和营养方面的耗费看来也不是过分的。所以,有可能大多数非基因DNA事实上就是废物或自私DNA。

然而,维持大量的非基因DNA也可能会有一些不利之处。首先,大基因组对诱变剂的敏感性表现得比小基因组的高(Heddle和Athanasiou,1975)。其次,维持和复制大量的非基因DNA可能会给该生物带来一定的负担,特别是当基因组的绝大部分都是非基因的时。

卡瓦利耶—史密斯(Cavalier— Smith,1978,1985)争辩说,一定有一种维持大基因组的"主要的进化力量"。他的假说是,这种DNA起着决定细胞核体积的核骨架作用。因为细胞越大要求细胞核也越大,所以,导致某种特别的细胞体积的选择,将会次生出导致某种特别的基因组大小的选择的结果。根据这一蓝图,过剩DNA是由选择来维持的,而它的核苷酸组成则可随机地改变。非基因DNA的附加的核类型功能可能是机械的。例如,一种高度重复、串接排列的卫星DNA,可能会影响染色体建筑样式,特别是它的曲度和核粒的相位。有些卫星DNA已知还与特异性的染色体蛋白结合,后者则可能在减数分裂和有丝分裂期间影响染色体的凝聚,因而也许有基因调控方面的效应(Levinger和Varshavsky,1982;James和Elgin,1986)。

没有一种解释看来是已解决了C—值矛盾问题的。以上所有机制,以及许多另外的机制,都可能只对维持"过度的"基因组大小有贡献,而我们将来的任务则是决定每种机制的相对贡献。

8.7 细菌中的GC含量

在真细菌中,基因组序列中鸟嘌呤和胞嘧啶的平均百分比,或GC含量(GC content),在从约25%到75%范围内变化。在许多情况下,细菌的GC含量看来与系统发育有关,亲缘关系接近的细菌有着类似的GC含量(图8—4)。

解释细菌中GC含量的变异本质上有两类假说。选择论者把GC含量看成是对环境条件的一种

适应形式。例如，生活在非常热的生态位中的嗜热细菌，强烈地倾向于应用由富 GC 的密码子编码的对热稳定的氨基酸（例如丙氨酸和精氨酸），且极力避免用由少 GC 的密码子编码的对热不稳定的氨基酸（例如丝氨酸和赖氨酸），这一现象已有若干报导（例如，Argos 等，1979；Kagawa 等，1984；Kushiro 等，1987）。因此，GC 含量可能是一种由选择决定的性状。另一个涉及选择论的事件要借助紫外线辐射作为选择力量。因为 T－T 二聚体对辐射敏感，所以，土壤上层的微生物，由于它们常暴露在阳光下，就应该比不暴露在阳光下的，比如说象 *E. coli* 那样的肠道细菌，含有更高的 GC 含量（Singer 和 Ames，1970）。

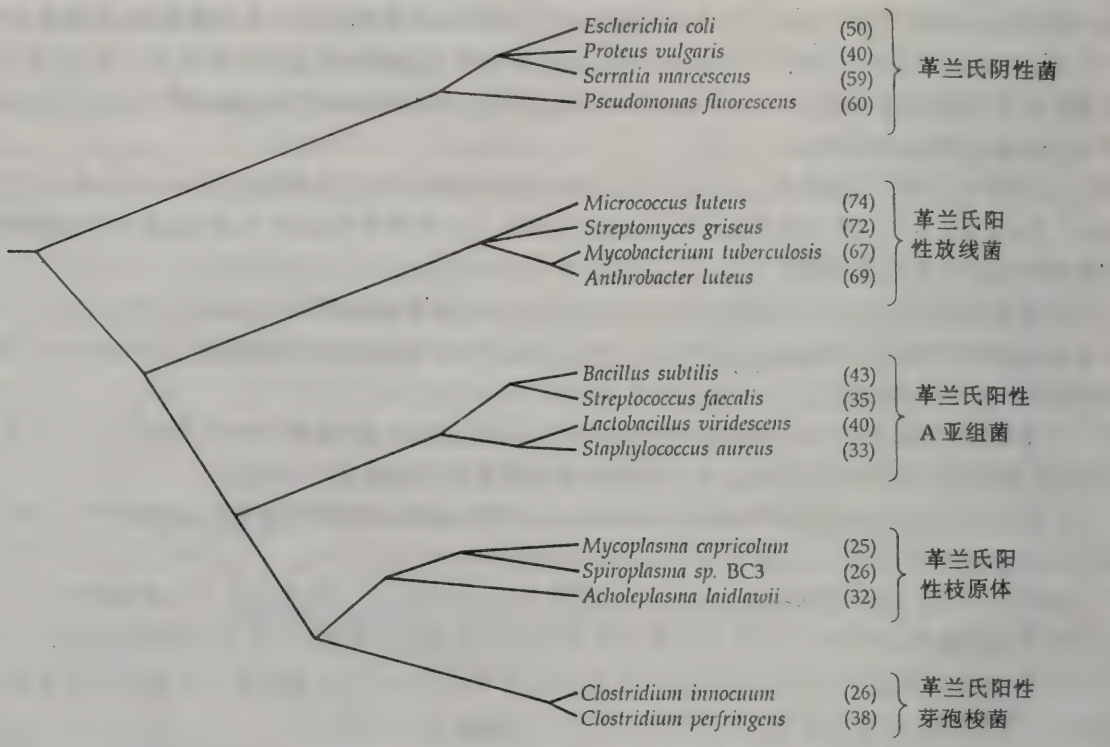


图8－4　几种真细菌中基因组的 GC 含量（括号中的数字）。该系统树是根据5S rRNA 序列构建的。分枝是无尺度的。自 Hori 和 Osawa(1986)，和 Muto 等(1986，1987)的资料结合而成。

突变论者的观点则以突变模式中的倾向性来解释 GC 含量上的变异（Suevka，1964；Muto 和 Osawa，1987）。根据这一观点，某一给定细菌种的 GC 含量是由以下两个因素的平衡所决定的：(1)从 G 或 C 到 T 或 A 的替换速率，用 u 表示，和(2)从 A 和 T 到 G 或 C 的替换速率，用 v 表示。处于平衡时，GC 含量预期为

$$P_{GC} = \frac{v}{v+u} \tag{8.1}$$

因此，

$$\frac{u}{v} = \frac{1-P_{GC}}{P_{GC}} \tag{8.2}$$

比率 u/v 又称 **GC 突变压力**（GC mutational pressure）。当 u/v 为3.0时，平衡点处的 GC 含量将为25%。枝原体 *Mycoplasma capricolum* 中的情形就是如此。当该比率为1时，GC 含量将为50%，象在 *E. coli* 中那样。当它为0.33时，GC 含量将为75%，如在 *Micrococcus luteus* 所示。不过，要估计 GC 突变压力，最好选用没有选择限制的部位而不要用总 GC 含量。例如，*Mycoplasma capricolum* 的为蛋白质编码基因中，四重简并位点上的 GC 含量低于10%（图8－5）。因而，u/v 一定高于9。类似地，*Micrococcus luteus* 中简并位点上的 $P_{GC}>0.9$，所以 $u/v<0.11$。

除 GC 突变压力外，突变引起的变化也受选择限制。换句话说，替换的模式是由突变模式和对抗某些突变的纯洁化选择的模式所决定的（第四章）。某一特定区域中的选择限制越弱，GC 突变压力对 GC 组成的影响就越强。图8－5展示了总 GC 含量与3个密码子位置上 GC 含量间的相关，数据来自11

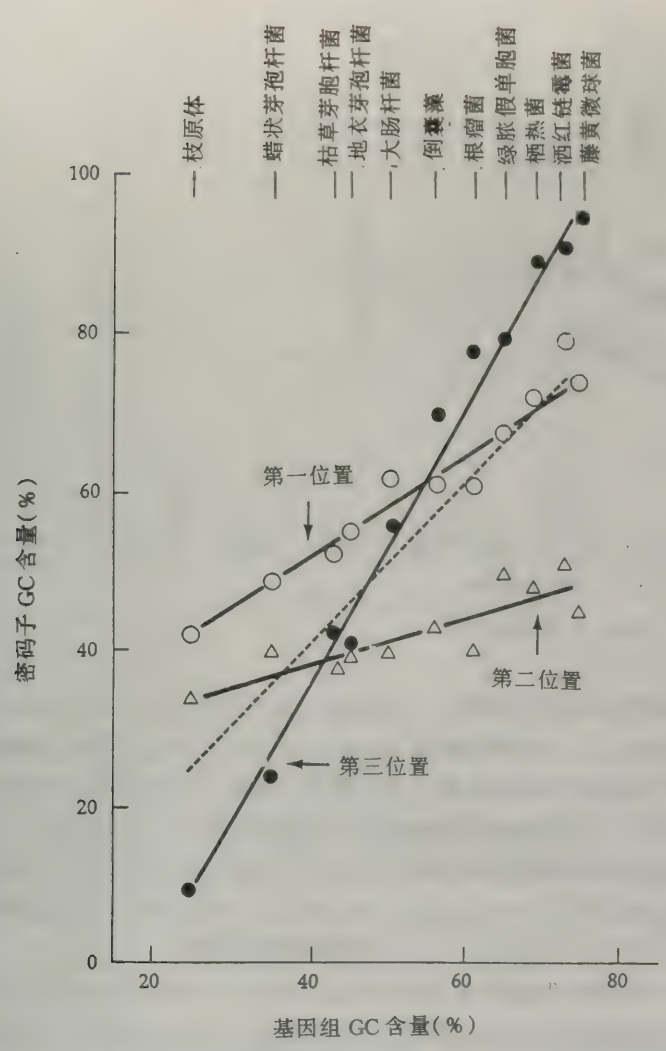


图8—5 总基因组 DNA 和第一、第二以及第三密码子位置间 GC 含量的相关。虚线表示基因组中与密码子中 GC 含量间完全对应时的理论期望关系。自 Muto 和 Osawa (1987)。

种细菌，覆盖了范围相当广的 GC 含量值。我们看到，第三密码子位置上的相关与无选择情况下的预期相关类似。另一方面，第一和第二密码子位置上的相关，在正相关的情况下，则表现出平缓得多的坡度。根据下面所说的事实这很容易得到解释，即作用在大多数简并的密码子第三位置上的选择，远不如第一和第二位上的那样严峻(第四章)，以致于第三位上的 GC 水平很大程度上是由突变压力来决定的。

8.8 脊椎动物基因组的组成上的组织化

图8—6展示了不同生物类群中的 GC 含量。有趣的是，虽然多细胞的真核生物的基因组大小一般要比原核生物的大，但真核生物中的 GC 含量却表现出了小得多的变异。特别地，脊椎动物的基因组显示出有相当一致的 GC 含量，处在从约40%到45%的范围中(Sueoka，1964)。脊椎动物中的 GC 含量处于一个较小的范围内的部分原因也许是，脊椎动物与细菌不同，它们相互分歧的时间还不够长，所以还不能在 GC 含量上积累较大的差异。

虽然脊椎动物在基因组的 GC 含量上有一致性，但它的基因组却有着比原核生物基因组更复杂的组成上的组织化。当脊椎动物的基因组被随机地裁剪成长30--100kb 的片段，并且将这些片段按其碱基组成分离时，这些片段即聚类成为数不多的、可通过其 GC 含量而相互区别的(富 GC 片段比富 AT 片段重)类型。每一类都以有虽不等同但却相似的碱基组成带为特征(Bernardi 等，1985；Bernardi，

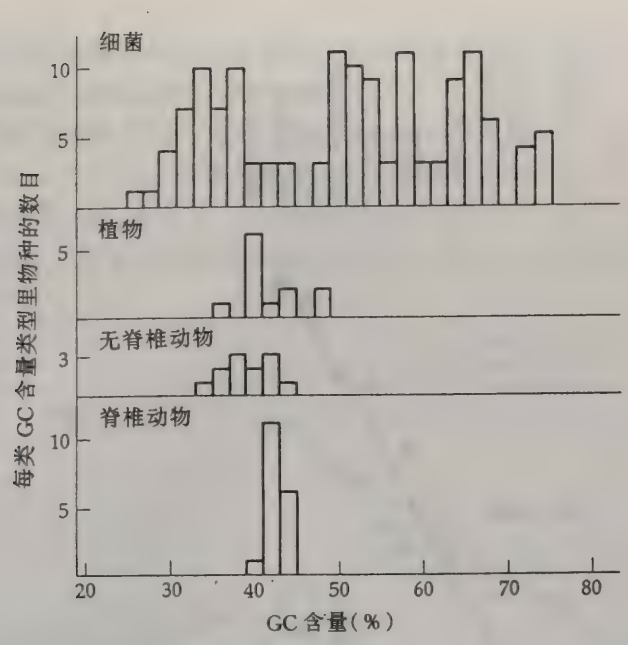


图8－6　不同生物类群中的 GC 组成。自 Sueoka(1964)。

1989)。

温血和冷血脊椎动物的基因组间在组成上的组织化方面有明显差异(Bernardi 等,1985,1988)。图8－7a 展示了,来自鲤 *Cyprinus carpio* 和两栖类 *Xenopus laevis*(左图),以及来自3种温血脊椎动物:鸡、小鼠和人(右图)的主要 DNA 组分的相对量和浮力密度。在鸡、小鼠和人的基因组中,有两个轻组分(L_1和L_2),代表了约三分之二的基因组,以及两个或三个重组分(H_1,H_2,和H_3),代表余下的三分之一。相比之下,来自大多数冷血脊椎动物的基因组 DNA 则主要为轻组分(图8－7a)。例如,在 *Xenopus* 中,密度高于1.704g/cm³的 DNA 片段占基因组的比例少于10%,相比之下温血脊椎动物则占30－40%。

DNA 片段的组成上的分布很大程度上是与这类片段的大小独立的,这指示出在非常长的 DNA 链中有某种组成上的同质性。这种同质的长段术语称为同质段(isochore)。图8－7b 展示了来自温血脊椎动物的细胞核 DNA 的镶嵌式组织化(即,轻与重同质段轮流出现)。当这些同质段在 DNA 裁剪期间断裂时,4种较大的有不同 GC 含量的分子家族即产生了。贝尔纳迪等(Bernardi 等,1985)的结论是,富 GC 的(重的)同质段约占温血脊椎动物基因组的三分之一,而在冷血脊椎动物中则几乎没有。

温血脊椎动物的基因组是镶嵌式的,这一发现与染色体分带研究的结果是一致的。温血脊椎动物的中期染色体,当用萤光染料、蛋白酶水解,或有差别的变性条件处理时,将显现出清晰的吉姆萨(Giemsa)暗带(G－带)和亮带(R－带)。与之成对照的是,冷血脊椎动物的中期染色体则只显出很少的分带,或完全不分带。因此,曾有建议认为,少 GC 和富 GC 同质段大致上分别对应于 G－带和 R－带(Comings,1978;Cuny 等,1981)。关于基因的复制时序的研究表明,位于富 GC 同质段(R－带)中的基因在细胞周期的早期复制,而位于少 GC 同质段(G－带)中的基因则在晚期复制(Goldman 等,1984;Bernardi 等,1985;Bernardi,1989)。

同质段中基因的位置

很多来自人和其他温血脊椎动物的基因组中的基因,曾用与适当的探针杂交的方法而被定位在这类组成上的同质段上(Bernardi 等,1985)。这些研究指出,基因在整个基因组中的分布是明显地非随机的,大多数基因位于仅占基因组的3－5%的最重组分(H3)中。

在大多数情况下,基因是处在有与其自身类似的 GC 含量的 DNA 片段中。这一发现为同质段的存在提供了独立的证据,同时也为同质段有较大的尺寸提供了独立的证据。事实上,因为构成 DNA 制备物的这些片段是经随机降解而产生的,所以,这些带有基因的片段在组成上处于较狭窄的范围中的现象表明,它们在约两倍于这些片段自身大小的区域内(即高到200kb),有着非常同质的碱基组

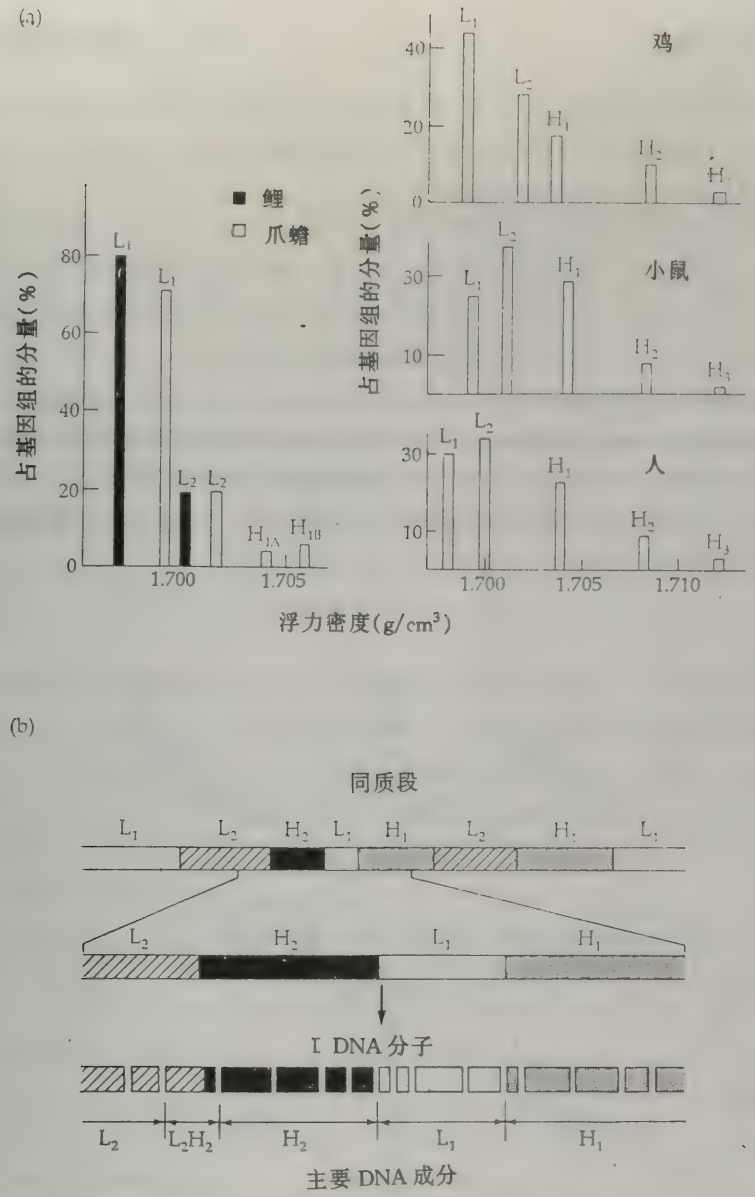


图8—7 (a)直方图展示，来自 *Cyprinus carpio*(鲤)和 *Xenopus laevis*(爪蟾)(左图)和来自鸡、小鼠与人(右图)的主要DNA同质段的相对量和浮力密度。(b)描绘温血脊椎动物细胞核DNA的镶嵌组织化的模式图。当这些同质段在DNA制备期间随机断裂时，4种较大的具不同GC含量的分子家族即产生了。还有几种较小的杂种家族产生。自Bernardi等(1985)。

成。

DNA序列数据分析已揭示出，在基因、外显子和内含子的GC水平，与它们处于其中的大段DNA区域的GC水平间存在正相关(Bernardi和Bernardi，1985；Bernardi等，1985；Ikemura，1985；Aota和Ikemura，1986)。图8—8将人类中的α—和β—珠蛋白簇进行了对比。β和类β珠蛋白基因是低GC含量的，且它们处在低GC区域中。反之，α和类α珠蛋白基因是富含GC的，则它们处在富含GC的区域中。同样的情形也曾在兔、山羊和小鼠中发现。在鸡中，β—和α—珠蛋白基因都是富含GC的，而且两者都处在富GC的区域中。相比之下，*Xenopus*(爪蟾)中的α—和β—珠蛋白基因则是少GC的，并且两者都处在少GC的区域中。

在绝大多数情况下，编码区中的GC含量有高于侧区中含量的倾向(图8—9)。我们还看到，第三密码子位置上的GC水平，平均下来高于内含子中的水平，而后者则又比5′和3′侧区域中的水平高。5′侧区中的GC水平又高于3′侧区中的，这或许是因为启动子及其周围区域倾向于富含GC的缘故。

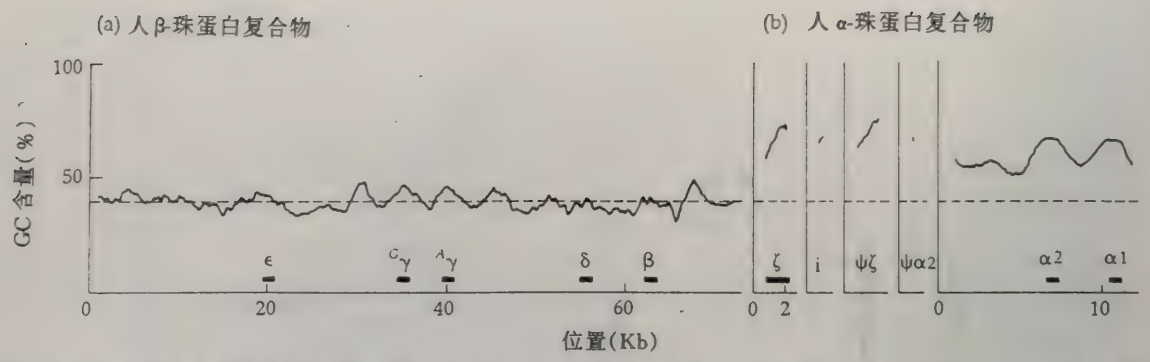


图8 8 GC 含量沿人的珠蛋白 DNA 顺序的分布：(a)β－珠蛋白基因簇；(b)α－珠蛋白基因族(不完全)。这些基因(粗线)以与图6－7中同样的次序排列。这些基因的名称显示在该图的底部；区域 i 是 ζ 和 ψζ 之间的基因间区。在 β－珠蛋白簇和覆盖 α1－和 α2－珠蛋白基因的区域中，每一点代表围绕该点的2001个核苷酸的 GC 组成的平均值，而其他区域中的每一点则代表1401个核苷酸的平均值。水平的虚线表示整个人的基因组的 GC 含量(40%)。自 Ikemura 和 Aota(1988)修改而成。

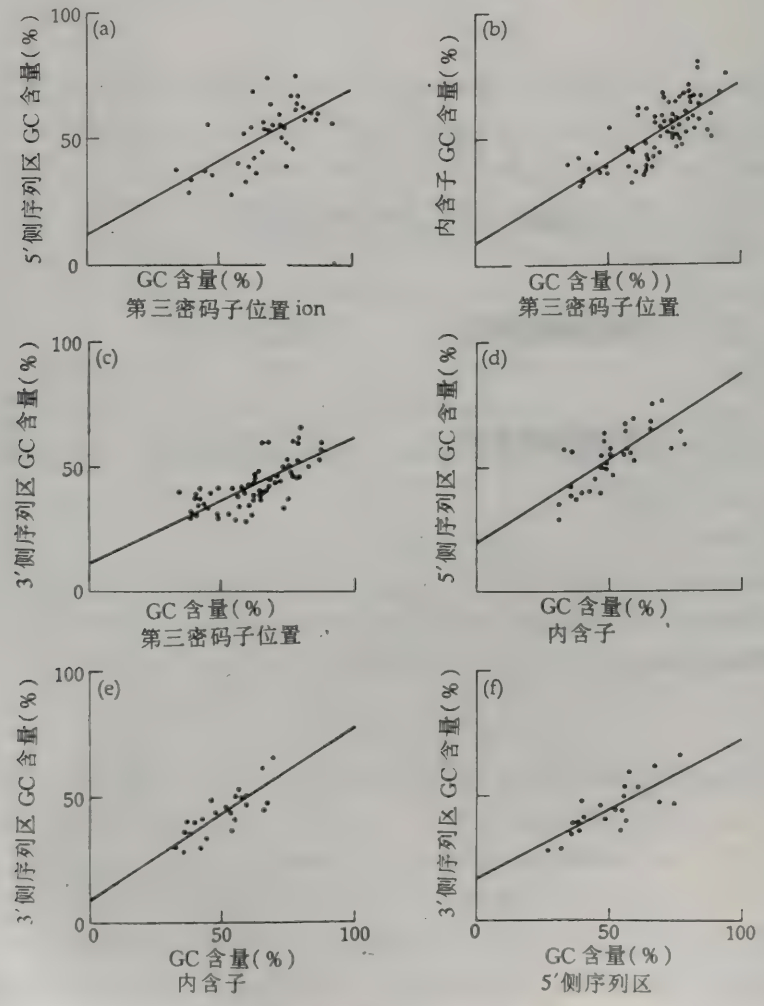


图8－9 基因的各个区域中的 GC 百分比含量间的关系。(a)第三密码子位置与5′侧区。(b)第三密码子位置与内含子。(c)第三密码子位置与3′侧区。(d)内含子与5′侧区。(e)内含子与3′侧区。(f)5′侧区与3′侧区。自 Aota 和 Ikemura (1986)。

同质段的起源

同质段的起源一如其被争论的那样神秘莫测。注意，问题在于 DNA 的长区段（30kb 或更长一些），它的普遍趋势是或者是富 GC 的或者是少 GC 的，而不是在 GC 含量上发生局部变异一象我们在基因的各不同区域中所看到的那样。贝尔纳迪等（Bernardi 等，1985；1988）提出，同质段是因功能上的（即选择上的）优势而产生的。他们的主要论据是，在温血生物中，GC 含量上的增加可以保护 DNA、RNA 和蛋白质免受热的降解（见下面），因为 G－C 键是比 A－T 键强的化学键（第一章）。我们称这个观点为**选择论者的假说**（selectionist hypothesis）。

沃尔夫等（Wolfe 等，1989a）提出，同质段是种系 DNA 复制期间，由于前体核苷酸库中的组成变化而造成的突变倾斜所引起的。富含 GC 的同质段由在种系细胞周期的早期复制的 DNA 区域所携带，而在那一期间前体库有较高的 GC 含量，因而就有一种突变成 GC 的偏向。反之，富含 AT 的同质段在细胞周期的晚期复制，那时前体库有较高的 AT 含量，故而有一种突变成 AT 的倾向。我们称它为**突变论者的假说**（mutationist hypothesis）。此假说是根据这样的观察而作出的：核苷酸前体库的组成在细胞周期内会发生变化，而这些变化事实上能导致新合成的 DNA 中碱基比改变（Leeds 等，1985）。必须清楚，哺乳动物基因组的复制是一个相当漫长的过程，要花8小时或者更多时间（Holmquist，1987）。

一个支持选择论者假说的理由是，第一和第二密码子位置上 GC 含量的增加可给予蛋白质以热稳定性，而内含子，第三密码子位置和不翻译区中 GC 含量的增加，则既能增加原始 mRNA 转录本的热稳定性，又能稳定染色体的结构，后者或许是通过影响 DNA－蛋白质的相互作用而实现的。事实上，嗜热细菌强烈地倾向于采用富 GC 密码子的情况已有报导（见第131页）。然而，温血脊椎动物的体温要比嗜热细菌所经受的温度低得多，所以，对脊椎动物的蛋白质和 DNA 序列的进化来说，温度可能并不是一个非常重要的因素。

选择论者假说所面临的一个困难是，大部分哺乳动物和鸟类的基因是低 GC 含量的这样一个事实。此假说也不能解释为什么有些重复基因有着相反的 GC 含量。例如，在哺乳类中，β－珠蛋白簇是低 GC 的，而 α－珠蛋白簇则是富 GC 的（图8－8），尽管这两类珠蛋白基因是在同一细胞内、同一时间里表达的，且有着同样的功能。类似地，有些免疫球蛋白基因位于富 GC 区，而另一些则位于富 AT 区（见 Aota 和 Ikemura，1986）。贝尔纳迪等（Bernardi 等，1985，1988）的解释是，同质段代表选择的单位，而哺乳动物中的 α 簇已被易位到了富 GC 同质段，β 簇则被留在低 GC 同质段中。根据这一理由，鸡中的 α 和 β 簇应该都被易位到富 GC 同质段了。然而，这一理由却引出了一个富 GC 同质段的功能优势问题。如果这一优势不是来自它所含有的基因，那么它来自何方呢？

突变论者的假说能解释哺乳类基因组中 α－和 β－珠蛋白簇间 GC 含量上有较大差异的成因，即假定它们分别位于早期和晚期复制的区域内。不过，它也面临着一些困难（Bernardi 等，1988）。例如，结构异染色质，如卫星 DNA，大多数是富 GC 的；而功能异染色质，如非活性的 X 染色体，在细胞周期的尾声阶段复制；在这些情况下，核苷酸库和 DNA 组成上的变化间看不到有什么联系。

结论是，现有的资料看来还不足以对这两种假说加以评判。也有可能突变压力和自然选择两者，都在形成温血脊椎动物基因组的组成上的组织化中起着作用。

习题

1、表8－5列出了4种原生动物的基因数，平均基因大小的大致估值和 C 值。（a）假定基因组的基因部分仅由为蛋白质编码的基因构成，那么，每一种生物中非基因 DNA 的比例是多少？（b）基因组大小和非基因 DNA 的比例间存在某种关系吗？（注意 C 值以 pg－微微克为单位）。

表8－5　4种原生动物的前体 mRNA 的平均大小和数目的大致估值和 C 值

物种	基因组大小(pg)	基因大小(核苷酸)	基因数
Physarum polycephalum	0.57	1500	20000
Oxytricha nova	0.4	2200	24000
Euplotes aediculatus	0.3	1800	40000
Dictyostelium discoideum	0.036	1500	6500

资料取自 Cavalier-Smith(1985)。

2、表8－6列出了4种生物的前体 mRNA 和成熟 mRNA 的平均大小的大致估值和 C 值。C 值上的差异，能用这些生物的基因间在非编码区的大小上存在差异来加以解释吗？(注意 C 值以 pg 为单位给出)。

表8－6　4种动物的平均基因大小和 mRNA 大小的大致估值和 C 值

属	基因组大小(pg)	前体 mRNA 大小(核苷酸)	mRNA 大小(核苷酸)
Drosophila(果蝇)	0.18	4200	2100
Aedes(蚊子)	0.83[a]	8400	2100
Strongylocentrotus(海胆)	0.89	8800	2100
Homo(人)	3.50	10000	2100

资料取自 Cavalier-Smith(1985)。

a、不同物种平均后得到的值。

3、假定在某一生物中，整个基因组都是由为蛋白质编码的基因所构成，20种氨基酸以相等的频率使用，(a)如果同义密码子的选取由尽可能导至最高 GC 含量的选择所决定，那么这种生物的 GC 含量将是多少？(b)如果同义密码子的选取由尽可能导致最低 GC 含量的选择所决定，那么这种生物的 GC 含量又是多少。

4、假定在某一 DNA 序列中，每核苷位点每世代的替换速率从 G 或 C 到 T 或 A 为 u，从 A 或 T 到 G 或 C 为 v。设 Pt 是世代 t 时 G 和 C 核苷酸在该序列中的比例。那么，GC 在下一世代中的比例为

$$P_{t+1} = (1-u)P_t + v(1-P_t)$$

根据这一等式，证明平衡点 GC 比例由等式8.1给出。

5、如果某一序列的起始 GC 含量为 P_0，则可证明世代 t 时的 GC 比例由下式给出：

$$P_t = \frac{u}{u+v} + (P_0 - \frac{v}{u+v})e^{-(u+v)t}$$

假定 $u=3\times10^{-8}$，$v=5\times10^{-8}$，且 $P_0=0.20$，计算 $t=10^7$时的 Pt 值。在这些条件下，GC 含量上的变化是一个缓慢过程还是一个快速过程？

后继阅读文献

Bernardi, G., B. Olofsson, J. Filipski, M. Zerial, J. Salinas, G. Cuny, M. Meunier-Rotival and F. Rodier. 1985. The Mosaic genome of warm-blooded vertebrates. Science 228: 953－958.

Blake, R. D. and S. Early. 1986. Distribution and Evolution of sequence characteristics in the *E. coli* genome. J. Biomol. Struct. Dynamics 4: 291－307.

Britten, R. J. and D. E. Kohne. 1968. Repeated sequences in DNA. Science 161: 529－540.

Brutlag, D. L. 1980. Molecular arrangement and evolution of heterochromatic DNA. Annu. Rev. Genet. 14: 121－144.

Cavalier-Smith, T. (ed). 1985. *The Evolution of Genome Size*. Wiley, New york.

Cold Spring Habor Symposia on Quantitative Biology. 1986. *Moleculur Biology of Homo sapiens*. Vol. 51. Cold Spring Habor Laboratory, Cold Spring Habor, NY.

Deininger, P. L. and G. R. Daniels. 1986. The recent evolution of mammalian repetitive DNA elements. Trends Genet. 2: 76－80

Dover. G. A. and R. B. Flavell (eds.). 1982. *Genome Evolution*. Academic Press. New York.

Jelinek. W. R. and C. W. Schmid. 1982. Repetitive sequences in eukaryotic DNA and their expression. Annu. Rer. Biochem. 51: 813—844.

John. B. and G. Miklos. 1988. *The Eukaryotic Genome in Development and Evolution*. Allen & Unwin. London.

Singer. M. F. 1982. Highly repeated sequences in mammalian genomes. Int. Rev. Cytol. 76: 67—112.

习题答案

第二章

2. $\Delta q=\dfrac{p^2q(w_{22}-w_{11})}{p^2(w_{11}-w_{22})+w_{22}}$

4. (a)$P_0=9.776\times10^{-4}$;(b)$P_5=0.246$;(c)$P_{10}=9.776\times10^{-4}$

5. $N_e/N=8/9$

6. $N_e=59.701$

7. $P=3.769\times10^{-4}$

8. (a)1-s 和 2-s 的编码区等同,因此,$x_i=2/3$,$x_j=1/3$,且 $\pi=1.729\times10^{-3}$;(b)$\pi=3.891\times10^{-3}$

第三章

6. (a)$P_s=0.615$;(b)$P_A=0.244$。

9. 当 $r=4$ 时,(a)$K=0.061$ 和(b)$K=0.056$

当 $r=6$ 时,(a)$K=0.041$ 和(b)$K=0.037$

第四章

2. 每位点 $K_I=0.012$,$K_{II}=0.048$ 和 $K_{III}=0.295$ 替换

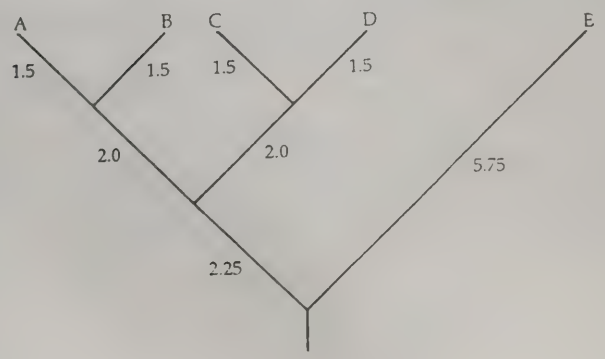
3. (a)每位点 $K=0.076$ 替换;(b)每位点 $K=0.075$ 替换。

4. 小鼠序列的进化比大鼠序列的快 1.63 倍。

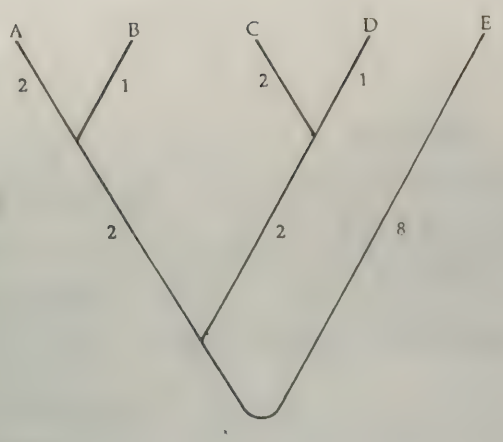
5. (a)每年每位点 $r_{max}=0.057$ 替换;(b)每年每位点 $r_{min}=0.004$ 替换。

第五章

3. (a)

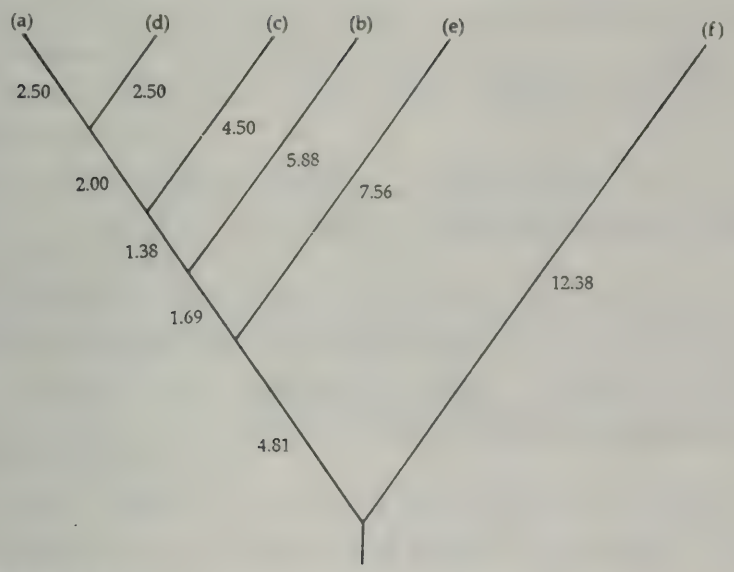


(b)



(c)与(b)相同。

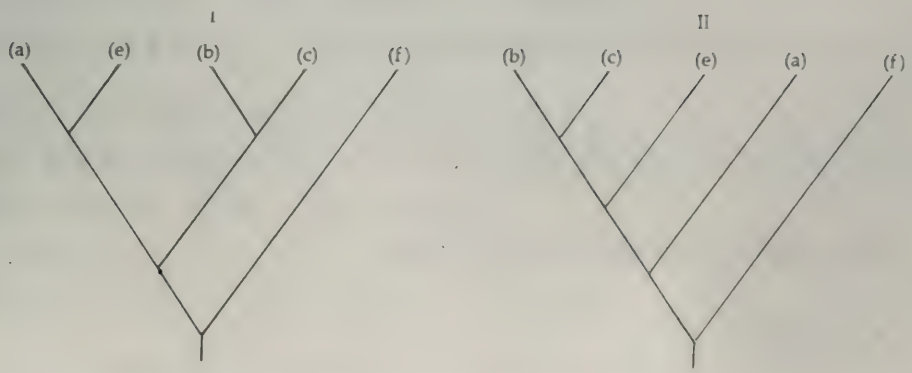
5. (a)



注意两个鳞翅目的物种，

(c)和(d)不形成一个进化枝。

(b)得到两个同样节省的树：



树Ⅱ中同翅目物种(b)、(c)和(e)形成一个进化枝，所以它与经典分类学结果一致。

第六章

2. (a)47 个非简并位点和 13 个四重简并位点；(b)70 个非简并位点且无四重简并位点；(c)所有

72 个位点都是非简并的。

3. (a)5;(b)6。

4. (b)17.24×10^6 年。

第八章

1. (a)*Physarum polycephalum*,5.37%;*Oxytricha nova*,13.46%;*Euplotes aediculatus*,24.48%;*Dictyostelium discoideum*,27.63%;(b)基因组大小与非基因 DNA 所占比例之间存在着负相关。

3. (a)56.76%;(b)30.00%

5. P_t=0.43

词汇表

A(1)在 DNA 或 RNA 中为腺嘌呤核苷,(2)在蛋白质为丙氨酸。

acceptor site 受体部位。内含子的 3′端。

active site 活性部位。一种蛋白质,通常为一种酶中的部位,其结构完整性为行使功能(例如基质结合)时所必不可少的。

additive tree 加性的树。一种系统树,其中任何两个终端节点间的距离等于将它们联结起来的枝长之和。

advantageous mntation 有利突变。一种能增加其携带者的适合度的突变。

alignment 线性排比。为了找出缺失和插入的位置而将两同源序列进行配比似的排列。

allele(allelomorph)等位基因(等位基因型)。一个基因座位上可供选择的基因形式。

allele frequency(gene frequency)等位基因频率(基因频率)。群体中某一给定基因座位上某一等位基因拷贝所占的百分比。

allozyme(allelozyme)异型酶(等位酶)。一种酶的某种等位形式。

alternative splicing 选择性拼接。通过采用不同的受体和给体部位,而从一种前体-mRNA 序列产生出两种或多种 mRNA 分子。

amino acid 氨基酸。一种具有通式 R-CH(NH_2)COOH 的有机分子,既有碱性基团(NH_2)又有酸性基团(COOH),还有为各种氨基酸所特有的侧链基团(R)。蛋白质的亚单位和结构砖块。

amino terminal(N-terminal)氨基末端(N-末端)。一个多肽的 NH_2 端。

amplification 扩增。某一基因或 DNA 序列的拷贝数的大幅度增加,增加量超出了作为该生物单倍体基因组的特征所具有的量。

analogy 类似。由趋同进化、而不是由共同的进化祖先所导致的相似。

aneuploidy(chromosomal duplication)非整数倍重复(染色体重复)。存在着额外染色体,以至一个细胞的染色体组成不是单倍体组的整数倍。

anticodon 反密码子。tRNA 分子中的核苷酸三联体,在翻译期间通过互补碱基配对而与 mRNA 中的特定密码子结合,从而在翻译期间确定某一具体氨基酸的位置。

antiparallel 反向平行。DNA 双螺旋的两条互补链的相反取向。

apoprotein 脱辅基蛋白。一种没有为其功能所必需的辅酶、辅因子和辅基的蛋白质(例如载脂蛋白)。

archaebacteria 古细菌。细胞壁中不含胞壁酸的原核生物;是一个高度多样化的类群,包括产甲烷细菌,嗜热、嗜酸的菌和嗜盐(在高盐浓度中生活)菌,据推测是生物界 3 个原始的后代品系之一。

arithmetic mean 算术平均。n 项之和用 n 来除。

asymmetrical exon 不对称外显子。两侧由不同相位类型的内含子所夹拥的外显子。

autosome 常染色体。除性染色体外的任何一种染色体。

back (backward) mutation 回复突变。使某一核苷酸位点回复到以前状态的突变。

bacteriaphage(缩写为 phage)噬菌体。寄生于细菌中的病毒。

balanced polymorphism 平衡多态。一种长时期内稳定而由平衡选择来维持的多态。

balancing selection 平衡选择。其结果造成群体中某一基因座位上维持着两个或多个等位基因的选择体制(例如超显性)。

banding 分带。染色体上明和暗地染色区域。

base(见 nucleotide)碱基。

base pair 碱基对。(1)某一核酸中一条链上的核苷酸,按嘌呤与嘧啶间的配对规则与另一条链上的核苷酸形成氢键。(2)度量双链 DNA 长度的单位。

biased codon usage(见 unequal codon usage)倾向性的密码子应用。

bifurcation(dichotomy)两分叉(两分枝)。系统树中一次物种形成的进化事件的图形表示,藉此一个祖先分类单位分裂为二。

bottleneck 瓶颈。群体大小上的锐减。

box 块(框)。邻近或位于某一基因内的短 DNA 序列,起着某种调控作用(例如 TATA 块)。

branch 分枝。系统树中某种进化关系的图形表示。

C (1)在 DNA 或 RNA 中为胞嘧啶核苷,(2)在蛋白质中为半胱氨酸。

C—terminal(见 carboxy terminal) C—末端。

C value(genome size)C 值(基因组大小)。某一物种的单倍体基因组所具有的特征性 DNA 的量。

C—value paradox C-值矛盾(C-值悖论)。C—值和形态学复杂性水平间明显地缺乏相关性的现象。

capping 戴帽子。真核生物中前体 mRNA 的 5′端的修饰过程,在此过程中 GTP 经 5′-5′三磷酸键而加入到该分子之中。

cap site(transcription initiation site)帽子部位(转录起始部位)。在 DNA 中是转录开始的部位,在 RNA 中是 mRNA 成熟过程期间被加上帽子的部位。

carboxy terminal(C-tenninal)羧基末端(C—末端)。一个多肽的 COOH 端。

carrying capacity 负载力、容纳量。在一个有限的栖息地中,某一给定物种所能维持的最大个体数。

catalyst 催化剂。一种能降低某一化学反应所需要的活化能,却不被该反应所消耗或改变的化合物。

cDNA(见 complementary DNA)

census population size(见 papulation size)统计记录的群体大小。

central dogma 中心法则。以 DNA 为遗传物质的生物中、信息流动的路线(DNA→RNA→蛋白质)。

chimeric protein(见 mosaic protein)嵌合蛋白质。

chloroplast 叶绿体。一种含叶绿素的、由膜包被的细胞器,它是植物和某些原生生物的细胞中进行光合作用的场所。

chromatid 染色单体。(1)由染色体复制而产生的两个拷贝之一。(2)构成该染色体的两个双链 DNA 分子之一。

chromosomal duplication(见 aneuploidy)染色体重复。

chromosome 染色体。在原核生物中指包含基因组的 DNA 分子。在真核生物中,指线性 DNA 分子与蛋白质复合而形成的含有遗传信息的线状结构。

cis 顺。两个序列或基因位于同一染色体上的排列。

clade 进化枝。(1)按严格定义,即由一个物种及其所有代表着进化树上的一个单系分枝的后代所构成的某一分类单位。(2)若较不严格地应用,则在前面的范围中可将不代表该分枝的后代除外。(3)关于现存生物,即指在所考虑的一大群生物中,一个有共同祖先的亚群,该共同祖先不是此群中其他生物的祖先。

cladogenesis(见 speciation)分枝进化、系统发生。

cladogram 进化分枝图、进化树。描绘或力图描绘一些群体、物种,或更高的分类单位间的进化关系的图形表示。

classification(见 taxonomy)分类。

coding region 编码区(域)。一个为蛋白质编码的基因中,最终将被翻译的所有外显子部分。

codominance(genic selection)共显性(基因选择)。二倍体基因组中某一基因座位上的两个等位基因对适合度有等同的贡献。

codon 密码子。mRNA 中相邻核苷酸构成的三联体,为由某一特定 tRNA 携带的氨基酸编码,或者确定翻译过程的终止。

codon family 密码子族。所有为同一氨基酸编码的密码子,它们相互间仅在第三密码子位置上有差异(例如,为亮氨酸编码的 6 个密码子中,UUA 和 UUG 组成一个族,CUU、CUC、CUA 和 CUG 则组成另一族)。

codon usage 密码子应用。一个密码子族中的成员在为蛋白质编码的基因中被使用的频率。

coenzyme 辅酶。一种不与酶结合,但作为一种电子、原子或原子基团的中间载体而为该酶的功能所必需的非蛋白质有机分子。

cofactor 辅因子。为某种酶行使功能所必需的无机分子。

coincidental evolution(见 concerted evolution)并发进化。

coincidental substitution 并发替换。两同源序列中的同一核苷酸位点上发生的两个替换。

colinearity 线性对应。无内含子基因的 DNA 顺序和其所编码蛋白质的氨基酸顺序间的精确对应。

complementarity 互补(性)。双链 DNA、双链 RNA 或 DNA－RNA 双链中核苷酸的反向平行配对。

complementary DNA(cDNA)互补 DNA。以 RNA 为模板由反转录酶合成的 DNA。

complex(composite)transposon 复合(合成)转座子。两侧由两个完整而独立的可转座插入序列夹拥的转座子。

compositional assimilation 组成同化。假基因中由于点突变的积累最终抹掉了它与产生它的功能基因间的顺序相似性,从而使其核苷酸组成与其邻近 DNA 序列的相似。

concerted evolution(horizontal evolution,coincidental evolution)协同进化(水平进化、并发进化)。一个物种中某一基因家族的成员间核苷酸序列的同质性的维持,尽管核苷酸序列随着时间而改变。

conditional fixation time 条件固定时间。一个最终将会在群体中固定的突变型等位基因达到固定的时间。

consensus sequence 共同顺序。在许多同源序列中,每一个位点上都代表了占绝大多数的核苷酸或氨基酸的那种顺序。

conservative substitution 保守替换。某一氨基酸被另一个有相似的化学性质的氨基酸替换。

conservative transposition 保守(型)转座。可转座因子不复制而从一个基因组位置移到另一个位置的运行。

constant site or constant region 恒定位点(不变位点)或恒定区。在所有被比较的同源序列中,DNA 内由同样的核苷酸占据的位点或区域。

convergence 趋同。相似的遗传或表型性状的独立进化。

convergence substitution 趋同替换。在两同源序列中的同一核苷酸位点上,两个不同的核苷酸由同一核苷酸所替换。

crossing-over 交换。两同源染色体间导致连锁基因重组的遗传物质交换的过程。假定该过程是两染色体先在同源位点处断裂,然后交换、接着重新连结而完成。

cyanobacteria 蓝细菌。具有光合作用能力的一类光合真细菌。以往称蓝绿藻。

D 天冬氨酸。

Darwinian fitness(见 fitness)达尔文适合度。

decoding(见 translation)解码。

degenerate code 简并密码。一种遗传密码,其中有意义密码子的数目大于氨基酸总数,结果有些氨基酸将由一个以上的密码子来确定。所有已知的遗传密码都是简并的。

degenerate site 简并位点。一个密码子中能由一个以上的核苷酸占据而仍为同样的氨基酸编码的

那种核苷酸位点。

degree of divergence 岐化程度。两同源序列相互间的差异程度。

deleterious mutation 有害突变。降低其携带者的适合度的突变。

deletion 缺失。从某一 DNA 序列中移去了一个或多个碱基。

denaturation 变性。蛋白质的三级结构的丧失。有时被用作 DNA 熔解的同义词。

deoxyribonucleic acid(DNA)脱氧核糖核酸。核苷酸连结而成的一种大分子聚合物,其中糖基为脱氧核糖。通常是双链的。在所有真核生物和原核生物中,以及许多病毒中,它是遗传信息的携带者。

deterministic process 决定性过程。其结果能从有关起始条件的知识来精确地预测的过程。

diagnostic position(见 informative site)判定位置。

dichotomy(见 bifurcation)两分枝。

digestion 消化、水解。双链 DNA 被某一限制性内切核酸酶所切割。

diploid 两倍体。一套染色体,其中每种染色体都有两个拷贝。

directional selection 定向选择。按一种特别的方向,或者走向固定或者走向灭绝,来改变某种等位基因的频率的选择体制。

disjunction 去联结。减数分裂期间同源染色体的分离,或有丝分裂期间互补的染色单体的分裂。

distance(见 genetic distance)距离。

distance matrix 距离矩阵。被研究的类群中各分类单位间的遗传距离值组成的矩阵。

divergence 分岐。一个分类学单位分成两个的分枝。(还可见 sequence divergence)

DNA(见 deoxyribonucleic acid)

DNA-DNA hybridization DNA-DNA 杂交。形成异源的 DNA 双链。

domain(见 functional domain)域。

dominance 显性。在杂合子中某一等位基因显示其完全的表型效应的性质。

donor site 供体部位。一个内含子的 5′端。

dose repetition 剂量重复。某一 DNA 序列存在着多重拷贝,这可由某基因产物以相对于单拷贝序列而言呈增加量的形式产生来表明。

dot matrix 点(矩)阵。一种序列的线性排比方法,其中两序列分别写成矩阵的首列和首行、而点则置于有相同的列首和行首的矩阵元中。

downstream 下游。一个核酸上的某一参考点的 3′方向。转录推进的方向。

drift(见 random genetic drift)漂变。

duplex 双链、双螺旋。一个双链 DNA 或双链 RNA,或一个由单链 DNA 与 RNA 分子互补配对而形成的双螺旋。

duplication 重复,复制。基因组中某一 DNA 片段的两个拷贝存在或产生。

duplicative transposition(见 replicative transposition)复制型转座。

E 谷氨酸。

effective population size 有效群体大小。与随机遗传漂变有关的群体大小。群体中从事生殖的实际个体数。

electromorph 电泳型。由电泳移动性差异检出的蛋白质变异型(同功酶或异型酶)。

electrophoresis 电泳。将溶解的或胶体的颗粒在由场下根据它们的移动性分离的技术。电泳移动性有赖于该颗粒的大小、三维几何形状和电荷。

endosymbiosis 内共生。两种生物间的互利关系,其中一种生物、内共生者(endosymbiont),生活在另一种生物、宿主(host)的组织或细胞内。

endosymbiotic theory 内共生学说。该学说认为,自我复制的细胞器,如线粒体和叶绿体,原初本是自由生活的生物,后来进入有核的细胞并与之建立了共生关系,进而失去了独立生存的能力。

enzyme 酶。能催化特异化学反应的蛋白质或蛋白质复合物。

eubacteria 真细菌。细胞壁中掺入了胞壁酸的原核生物。即除古细菌外的所有细菌。生物界中三

个原始祖先品系之一。

eukaryote 真核生物。具有一个真正的细胞核和一些由膜包被的细胞器的生物。生物界中三个原始祖先品系之一。

exon 外显子。基因的一个 DNA 片段，其转录本在成熟的 RNA 分子中出现。

exon duplication 外显子重复。一个基因内的某一外显子的重复拷贝的产生。

exon insertion 外显子插入。一个或多个外显子从一个基因掺入到另一个基因子。

exon shuffling 外显子混匀。严格地讲指外显子重复和外显子插入。常常与外显子插入同义地应用。

expected heterozygosity(见 heterozygosity, gene diversity)期望杂合度、预期杂合度。

extinction 灭绝。一个进化谱系的终止。

F 苯丙氨酸。

fecundity 生殖力。一种适合度分量。某一给定基因型每个体的生育数或产卵量。

fertility 能育性。一种适合度分量。某一给定基因型每个体的成活后代数。

fitness(Darwinian fitness)适合度(达尔文适合度)。某一个体或某一基因型的生存和繁殖上的相对成就的测度。某一个体或某一基因型对将来世代的相对贡献。

fixation 固定。当某一等位基因在群体中的频率达到 100%时所出现的情况。一个两倍体群体中的所有成员对同一等位基因而言都是纯合的时的情况。

fixation probability 固定概率。某一特定等位基因将在群体中固定的可能性。

fixation time 固定时间。某一突变型等位基因在群体中达到固定所花的时间。

flanking sequence 侧(区)序列。被转录基因的 5′和 3′端处的不转录序列。

foldback DNA 自身折叠 DNA。含有完美或接近完美回文的 DNA，当其为单链时可通过自身折回而形成发夹状的结构。

fourfold degenerate site 四重简并位点。密码子中的一个核苷酸位点，在该位点上一切可能的替换都是同义的。

frameshift mutation 阅读框架移动突变。一种扰乱了为蛋白质编码基因的阅读框架的突变。一个 DNA 片段的缺失或插入，当该片段的长度不是 3 或 3 的倍数个核苷酸时，即可造成这类突变。

frameshifted protein 阅读框架移动后的蛋白质，由于阅读框架变得与基因的原初或主要阅读框架不同，而编码出一种完全或部分地不同的蛋白质。

functional constraint(selective constraint)功能限制(选择限制)。一个位点或一个基因座位对核苷酸替换所具有的特征性忍受程度。

functional domain(domain)功能域(域)。蛋白质内的一个界限明确的区域，能行使某一特殊的功能。它可能不是由连续的氨基酸段所构成，虽然就所涉及的蛋白质的三级结构而言，它几乎总是由那些相互邻近的氨基酸所构成。

G (1)在 DNA 或 RNA 中为鸟嘌呤，(2)在蛋白质中为甘氨酸。

gamete 配子。具单倍体数染色体的生殖细胞。

gap 裂缝。一段插入或缺失。在序列线性排比中则为一个含有空缺碱基的对子。

gap penalty 裂缝处罚。与点替换发生的频率相比裂缝事件在进化中发生的频度如何？对此问题作出的估价即裂缝处罚。在线性排比算法上，或者用一个因子来乘以裂缝的总长度，或者将某一给定长度裂缝的数目乘以一个函数，应用这些值才有可能对裂缝和替换加以比较。

gene 基因。基因组 DNA 或 RNA 中的一个序列，它是某一特别功能的根本。

gene conversion 基因转变。一种非相互重组过程，结果是一个序列变得与另一序列等同。

gene duplication 基因重复。广义地，指一个 DNA 序列的两个拷贝产生。狭义而言，指一个完整的基因序列的重复。

gene family(见 multigene family)基因家族。

gene frequency(见 allele frequency)基因频率。

gene pool 基因库。一个有性生殖的群体中的所有基因。

gene sharing 基因分享。一个基因在不重复且失去其原始功能的前提下，获得并维持第二种功能的现象。

gene substitution 基因替换。一种过程，藉此一个新的突变型等位基因在群体中达到固定。

gene tree 一种由来各每物种的一个或几个基因构成的系统树。

generation time 世代时间。两个连续世代之间的平均时间间隔。有时定义成双亲在生产其顺序处在中间的孩子时，所具有的平均年龄。

genetic code 遗传密码。一组将密码子翻译成氨基酸的规则。

genetic distance(distance)遗传距离(距离)。广义地，指个体、群体，或物种间遗传差异程度的几种测度中的任何一种。对分子进化而言，则是两同源 DNA 序列自分岐以来，它们间累积的每核苷酸位点的核苷酸替换数的测度。

genetic drift(见 random genetic drift)遗传漂变。

genetic polymorphism(见 polymorphism)遗传多态性现象。

genetic DNA 基因(的)DNA。基因组中含基因的部分。

genetic selection(见 codominance)基因选择。

genome 基因组。由一个细胞或个体所携带的整套遗传物质。

genome doubling(见 polyploidy)基因组加倍。

genome duplication(见 polyploidy)基因组重复。

genome size(见 value)基因组大小。

genomic compartmentalization 基因组的区域化。指细胞中独立地复制的基因组的存在。通常，指细胞器的基因组。

genotype 基因型。某一生物的特别的等位基因组成，包括那些不在表型水平上表现出来的等位基因。常常指被研究的一个或少数几个基因的等位基因组成。

geometric mean 几何平均(值)。n 个项相乘以后开 n 次方所得到的根。

germ-line cell 种系细胞，生殖细胞。精细胞或卵细胞，或一个它们的前身细胞。

H 组氨酸。

haploid 单倍体。具有一套不成对的染色体的细胞或生物。

haploid set 单倍体组。一个单倍体细胞或生物中的染色体。

haplotype 单倍型。一条染色体的特别的等位基因组成。常常指被研究的一个或几个连锁基因的等位基因组成。

Hardy-Weinberg equilibrium 哈迪－温伯格平衡。一个两倍体群体中基因型的频率等于有关等位基因的频率之积的局面。

heteroduplex 异源双链。两条链各来自不同个体的双链核酸分子。

heterogeneous nuclear RNA(heterogeneous RNA. heteronuclear RNA，hnRNA)核内不均-RNA(异质 RNA、异核 RNA、hnRNA)。细胞核中的 RNA 转录本，表示 rRNA、mRNA 和 tRNA 的前体和经加工的中间产物，以及成熟但未传送到细胞质中的 RNA 转录本。

heterosis(见 overdominance)杂种优势。

heterozygosity 杂合度、杂合性。对群体中遗传变异的一种测度，可用杂合子对所有基因座位的平均频率算出(**观察杂合度**，observed heterozygosity)，或者用杂合子在一个处于哈迪－温伯格平衡的群体中预期的平均频率算出(**期望杂合度**，expected heterozygosity，或**基因多样性**，gene diversity)。

heterozygote 杂合子。在被研究的基因座位上有不同等位基因的两倍体个体。

heterozygote advantage (见 overdominance)杂合子优势。

higher taxon 高级分类单位。种以上水平的分类单位。

highly repetitive DNA 高度重复的 DNA。由平均重复成百上千次的序列构成的基因组 DNA 的部分。

highly repetitire genes 高度重复的基因。在单倍体基因组中出现许多拷贝的有功能基因。

homoduplex 同源双链。一种双链 DNA，其两条互补的链来自同一个体。

homology 同源(性)。因有共同祖先或遗传相关而类似。

homozygote 纯合子。在一个或多个基因座位上有相同的等位基因的两倍体个体。

horizontal evolution(见 concerted evolution)水平进化

horizontal gene transfer 水平基因转移，横向基因转移。遗传信息从一个基因组向另一个基因组的转移，特别是不同物种间的转移。

hotspot of mutation 突变(的)热点。基因组 DNA 的一个片段，对自发的或某种特别诱变剂作用下的突变表现出较高的倾向性。

hybrid dysgenesis 杂种劣势。一组相关的异常症状，在某种相互作用的果蝇品系间的一种类型的杂种中能自发地诱生、但在相反类型的杂种中则不出现。

hybrid vigor(见 overdominance)杂种优势，杂种兴旺。

hydrogen bond 氢键。氢原子和一个电负性的原子、如氧、之间的弱的、非共价键。

hypervariable site or hypervariable region 高可变部位或高可变区域。展示出超过了种间变异性的一段 DNA 或蛋白质。如此大的变异性的维持，通常需要该基因座位经受某种形式的平衡选择，如超显性选择。

I 异亮氨酸。

independent assortment(Mendel's second law)独立分配(孟德尔第二定律)。在非连锁的基因座位中，一个座位上的等位基因的分离与另一个座位上的分离无关。

inferred tree 推测树、推论树。根据属于现存分类单位的经验资料作出的系统树。

informative site(diagnositic position)信息位点(判定位置)。用于从所有可能的系统树中选取最节省树的位点。在分子进化中，是其中至少有两种不同类型的核苷酸或氨基酸，且它们中的每一种至少在两个序列中出现的位点。

initiation codon 起始密码子。为蛋白质编码基因的阅读框架中的第一个密码子；通常，在真核生物中是 ATG 编码的甲硫氨酸，在原核生物中则编码甲酰甲硫氨酸。

in-phase overlapping 相位内重叠。两个或多个蛋白质按同一阅读框架翻译的现象。

insertion 插入。一个或多个核苷酸插入一个 DNA 序列中的突变。

insertion sequence 插入序列。一种除带有转座所必需的片段外不带任何遗传信息的可转座因子。

internal gene duplication(partial gene duplication)基因内重复(部分基因重复)。基因内的重复序列，由比整个基因序列小的有关重复所派生而来。

internal node 内部节点。系统树中代表某一祖先生物或基因的图形。

interon (intervening sequence)内含子(间隔序列)。一种可转录基因的 DNA 片段，其转录本在 RNA 成熟的过程中被除去，因而不出现在成熟的 RNA 分子中。它位于外显子之间。

invariant repetition 不变重复。在顺序上相互等同或几近等同的重复 DNA 片段的存在。

inversion 倒位。造成某一 DNA 片段极性颠倒的突变。

isoaccepting tRNA 同受体 tRNA，同功 tRNA。能携带同样氨基酸的不同类型 tRNA。

isochore 同质段。在碱基组成上同质的基因组 DNA 片段。

isozyme(isoenzyme)同功酶。有相同或接近相同的化学性质，但由不同基因座位所编码的一些不同型式的酶。

junk DNA 废物 DNA。基因组 DNA 中的无功能部分。

K 赖氨酸。

L 亮氨酸。

lagging strand 后随链。以不连续方式，从 5′到 3′沿背离复制叉的方向合成的 DNA 链。

leading strand 前导链。以连续的方式，从 5′到 3′沿朝向复制叉的方向合成的 DNA 链。

length abridgment 长度缩短。假基因在进化期间由于缺失超过插入而造成的逐渐缩短。

lethal mutation 致死突变。造成其纯合携带者死亡或不育的突变。

ligation 连接。将双链 DNA 中由缺口分开的两个相邻碱基联接、而形成的磷酸二酯键。用于联接的粘性末端由限制性内切核酸酶水解产生。连接过程由连接酶(ligase)催化。

LINE(Long INterspersed Element 的首字母缩略)长散在因子。一种散在的重复序列,典型的 LINE 长 5000 碱基对以上,在多细胞真核生物的基因组中其拷贝数为 10^4 或超过 10^4。又称长的散在重复序列。

linkage 连锁。两个或多个非等位的基因位于相互很接近的位置处,并倾向于一起遗传的现象。

localized repeated sequences 区域性重复序列。串接排列的重复序列,通常由短的单纯重复单位所构成(例如卫星 DNA)。

locus(复数 loci)基因座位。染色体上某一特定基因或 DNA 片段所处的位置。

lowly repetitive genes 低度重复基因。在单倍体基因组中仅以几个拷贝存在的基因。

M 苏氨酸。

match 匹配。在序列排列时,两序列的同源位置上存在同样的碱基的现象。

maturation 成熟。从前 mRNA 形成 mRNA。

maximum parsimony(parsimony)最节省(节省)。从所有可能的系统树中选出所需替换数最少的那种系统树,以它作为真实系统树。

meiosis(reduction division)减数分裂。在从两倍体细胞产生单倍体配子时采用的真核细胞分裂过程。减数分裂的特征是分裂后染色体数目减半,以保证每个配子有各对常染色的一个代表和一半性染色体。

meiotic drive(见 segregation distortion)减数分裂驱动。

melting 解链、(双链的)熔解。双链变性后成为单链的核苷酸。

Mendelian segregation(Mendel's first law,segregation)孟德尔(式)分离(孟德尔第一定律、分离)。杂合子中某一基因对的两个不同等位基因在减数分裂中分离,以同样比例产生两类配子,每类配子各带一个不同的等位基因。

Mendelian second law(见 independent assortment)孟德尔第二定律。

—mer 表示蛋白质中亚基数的后缀。前缀表示亚基的数目(例如,monomer,dimer,tetramer,multimer),或单元间的相似性(例如,homomer,heteromer),或既表示单元数目又表示单元的类型(例如,homotrimer,heteromultimer)。

messenger RNA(mRNA)信使 RNA。从某种初级 RNA 转录本加工而成 RNA 分子,用于合成由氨基酸组成的多肽的翻译中。

middle-repetitive DNA 中(等程)度重复 DNA。基因组 DNA 中,长度相对较长、重复数平均在几十次到几百次之间的那一部分序列。

migration 迁移。指群体遗传学中个体或基因在各群体间的移动。

mismatch 匹配错误。在顺序的线性排比中,两序列的某一同源位置上有不同碱基的现象。

missense mutation(见 nonsynonymous mutation)误义突变。

mitochondrion(复数 mitochondria)线粒体. 真核细胞中一种含 DNA 的细胞器,它用一种需氧的电子传递系统,把由食物分子分解所得到的化学能转换成 ATP 储能。

mitosis 有丝分裂。真核细胞的分裂模式,通过这种分裂能产生两个具有与亲本细胞同样染色体套的子细胞。

mobile element(见 transposable element)可(移)动因子。

moderately repetitive genes 适度重复基因。单倍体基因组中具有适度拷贝数的基因。

module(structural domain)组件(结构域)。球状蛋白质中的、某个结构上独立的、稳定的、且紧密的空间单位。该单位能与其他部分相区别,通常是由一段连续的氨基酸段所组成。

molecular clock 分子(时)钟。(1)指突变在某一给定基因组片段中积累的速率。(2)指假说,即就任一给定基因或 DNA 序列而言,突变在所有进化谱系中以接近恒定的速率积累,只要该基因或该

DNA 序列保留其原初功能。将该时钟用于所有基因和所有生物的推广尚有争议。

monomorphic 单态(性)的。一个群体中所有个体在某一基因座位上实际上有同样的等位基因,则称是单态的。

monophyletic 单源的。享有一个共同祖先的现象。

mortality 死亡率。适合度的一个分量。某一给定基因型的个体在达到某一年龄之前死亡的平均概率(例如,到平均生育年龄的死亡率)。

mosaic protein(chimeric protein)镶嵌蛋白质(嵌合蛋白质)。由一个有来自不同基因的区域的基因编码的蛋白质。也指通过遗传工程而得到的人工蛋白质。

mRNA(见 messenger RNA)信使 RNA。

multifurcation 多(重)分叉。包括 3 个以上的分类学单位的系统树中,等级不明的分枝的图象表示。偶尔,也是同时产生两个以上物种的物种形成事件的图象表示。

multigene family(gene family)多基因家族(基因家族)。由某一祖先基因重复而产生的一套基因,它们间的相似度大于 50%。它们相互间常常紧密连锁,具有相似或重叠的功能。

multiple substitution 多重替换。在某一 DNA 序列的同一核苷酸位点上相继发生两次或多次替换。

mutagen 诱变剂。使突变率增加的物理学或化学实体。

mutant 突变型。某一些基因的一种新变异型。

mutation 突变。使某一 DNA 序列变成不同于其原来状态的改变。

mutation rate 突变率。某一个体中,每单位时间里每核苷酸位点或每基因产生的突变数。

mutational bias 突变偏斜。四种核苷酸表现出有不同的突变倾向,或者产生某一核苷酸多于其他核苷酸这样结果的突变模式。常常是由各核苷酸不均等地积累所造成的。

N (1)在 DNA 或 RNA 中,表示某一未知的核苷酸。(2)在蛋白质中表示天冬酰胺。

natural selection(selection)自然选择(选择)。由于各个体或各基因型间适合度上的变异性,造成某一物种的不同成员的生殖差异,从而导致了等位基因频率随着时间的改变而改变。

negative selection 负选择(见 purifying selection)

neighboring taxa 近邻分类学单位(见 sister taxa)。

neutral mutation 中性突变。不改变该生物的适合度的突变。

neutral theory(neutral-mutation theory 或 neutral-mutation hypothesis)中性学说(中性突变学或中性突变假说)。认为分子水平上的进化主要是由突变输入和随机遗传漂变所决定,而不是由自然选择所决定。

node 节点。系统树中某一现存的或祖先的 OTU 的图象表示。

nondegenerate site 非简并位点。编码区中所有替换都是非同义替换的核苷酸位点。

nondisjunction 不分离。减数分裂期间同源染色体的分离失败。

nonfunctionalization(silencing)无功能化(沉默)。继丧失功能的突变之后,功能基因变成假基因的转变。

nongenic DNA 非基因 DNA。基因组中不含基因的部分。

nonsense codon 无意义密码子(见 termination codon)。

nonsense mutation 无(意)义突变。使有意义密码子变成终止密码子的突变。

nonsense strand 无义链。基因的一条链转录产生 RNA,另一条不产生 RNA 的链即无义链,其转录本的顺序与 RNA 的互补。

nonsynonymous substitution(missense substitution)非同义替换(误义替换)。使某一密码子变成为另一氨基酸编码的密码子的替换。

N-terminal N-(末)端(见 amino terminal)

nucleic acid 核酸。DNA 或 RNA。

nucleotide(base)核苷酸(碱基)。由一个含氮碱基、一个核糖,和一个磷酸基团组成的分子。构成核

酸的基本建筑砖块。

nucleotide diversity 核苷酸多样度。用于测度核酸序列的多态性的尺度。从某一群体随机选取的任何两个序列间的每位点平均核苷酸差异数。

nucleotide substitution 核苷酸替换。发生了一个核苷酸被另一核苷酸替换的一种突变。在进化中指一个核苷酸被另一个将在群体中固定的核苷酸替换。

nucleotypic 核类型的。指 DNA 序列的除作为遗传信息的携带者以外的某种功能(例如,起作为细胞核骨架的作用)。

nucleus(复数 nuclei)细胞核。真核生物中,一个由膜包被而含有染色体的细胞器。

observed heterozygosity 观察到的杂合度(见 heterozygosity)。

ontogeny 个体发育、个体发生。生物从合子到成体的发育中的事件序列。

open reading frame(ORF)开读框架。具有可翻译成蛋白质潜力的 DNA 序列。

operational taxonomic unit(OTU)操作中的分类单位。任一被研究的现存分类学单位。

operon 操纵子。一种遗传学单位或由一个或多个基因构成的基因簇,这些基因作为一个单位而被转录,而且以协调的方式表达。

ORF(见 open reading frame)。

organelle 细胞器。严格地说指真核生物细胞内由功能膜包被的结构(例如,细胞核,线粒体,叶绿体)。通常都把细胞核排除在外。

orthology 垂直相关。作为物种形成事件的结果而产生的顺序相似。

OTU(见 operational taxonomic unit)。

outgroup 组外单位。在一群物种中,与其他物种亲缘关系最疏远的一个或一组物种。该分类单位在其他分类单位相互分岐之前,已从这群分类单位中分岐出来了。

out-of-phase overlapping 相外重叠。同一 DNA 序列中以不同读码框架为两个或多个蛋白质编码。

overdominance(heterosis,heterozygote advantage,hybrid vigor)超显性(杂种优势、杂合子优势、杂种优势)。由杂合子比两种纯合子有更高适合度而产生的选择形式。

P 脯氨酸。

palindromic sequence 回文顺序。对两条互补链读出相同结果的 DNA 或 RNA 顺序(如 AATGCATT)。表现出关于某一中心轴点对称的 DNA 或 RNA 顺序。

parallel substitution 平行替换。在两个或更多谱系中独立地发生位于同一核苷酸位点上的相同突变。

paralogy 平行相关。重复的祖先基因的后代间的顺序相似。

pararetrovirus 拟反录病毒。含有一个为反转录酶编码的基因,但自身不能插入宿主染色体的病毒。

parsimony 节省。从字面上解释,即以最少的操作而达到目的。(见 maximum parsimony)

partial gene duplication(见 internal gene duplication)部分基因重复。

pattern of mutation 突变的模式。指相对频率,一种核苷酸以这种频率突变成另一种核苷酸。

pattern of substitution(substitution scheme)替换的模式(替换方式)。指相对频率,在进化过程中一种核苷酸或氨基酸以该频率变成为另一种。

PCR(见 polymerase chain reaction)

phage(见 bacteriophage)噬菌体。

phase class 相位类型。内含子所处的位置,相对于两邻近的为蛋白质编码的外显子的阅读框架的类型。

phase -0 intron 相位-0 内含子.位于两密码子之间的内含子。

phase-1 intron 相位－1 内含子。位于某一密码子的第一和第二核苷酸之间的内含子。

phase－2 intron 相位－2 内含子。位于某一密码子的第二和第三核苷酸之间的内含子。

phenogram 表型关系图。根据一些个体、物种、或更高级的分类单位之间的所有相似性，以描绘或试图描绘它们间的分类学关系的一种图形表示。

phenotype 表型。某种遗传控制性状的可观察特征。

phylogenetic tree 系统(发育)树。表示一群分类单位或一群基因的系统发育的图形。

phylogenetics 系统发育学。重建一群分类单位或一群基因的进化史的科学。

phylogeny 系统发育。一群分类单位或一群基因及其祖先的进化史。

plasmid 质粒。一种自治且自我复制的染色体外环状 DNA。

point mutation 点突变。仅涉及一个核苷酸位点的突变。通常为一次核苷酸替换。

polarity 极性。核酸从 5′到 3′方向被阅读的性质，其结果与按相反方向阅读的不同。

polyadenylation signal 多聚腺苷酸化信号。大多数真核生物 mRNA 分子上都有的一个区块，决定多聚腺苷酸位点的定位。

polyadenylation site(poly(A)-addition site)多聚腺苷酸化位点(多聚(A)添加位点)。真核生物中大多数 mRNA 分子的 3′端部位。在该位点处将被添上一个多聚 A 尾巴。

polygamy 多配偶制。一种交配体制，在这种体制中或者一个雄性与一个以上的雌性交配(一夫多妻制)，或者一个雌性与一个以上的雄性交配(一妻多夫制)。

polymerase chain reaction(PCR)多聚酶链式反应。从未纯化的混合物中选定 DNA 序列加以扩增的方法。

polymorphism(genetic polymorphism)多态性(遗传多态性)。某一基因座位上两个或多个等位基因共存的现象。

polypeptide 多肽。由氨基酸通过肽键而相互共价连接而成的分子。通常，它用来指某一蛋白质的功能性三维构型确定前的氨基酸链。

polyphyletic 多源(发生)的。从不同祖先传下来的。

polyploidy(genome doubling，genome duplication)多倍体，多倍性重复(基因组加倍，基因组重复)。一个细胞或一个个体中存在多于两个单倍染色体组的现象(例如，四倍体、六倍体)。

polyprotein 多蛋白。一种翻译后分裂成两个或多个蛋白质的多肽。

population 群体。某一物种中共有一个基因库的一群个体。

population size(census population size)群体大小(统计群体大小)。一个群体的个体数。

positive selection 正选择。对某一有利突变等位基因的选择。

pre-messenger RNA(pre-mRNA)前体信使 RNA。某一为蛋白质编码基因的初级转录本，处于成熟前状态。

preproprotein 前蛋白原。处于任何翻译后变化发生之前的初级翻译产物。

pretermination codon 前终止密码子。只需一次突变即可变成终止密码子的密码子。

primary amino acid 基本氨基酸。由普适遗传密码决定的 20 种氨基酸中的任何一种。

primary structure 一级结构。多肽链中氨基酸的顺序。DNA 或 RNA 分子中核苷酸的顺序。

processed gene(见 retrogene)加工后基因。

processed pseudogene(见 retropseudogene)加工后假基因。

processed sequence(见 retrosequence)加工后序列。

prokaryote(bacterium)原核生物(细菌)。一种缺乏核膜、与 DNA 结合的组蛋白、和细胞器的生物。包括真细菌和古细菌。

proprotein 蛋白原。一种信号肽被去除而附加的翻译后修饰还未进行的翻译产物。

prosthetic group 辅基。与脱辅基蛋白结合以满足功能性要求的非蛋白分子(例如血红蛋白中的血红素)。

protein 蛋白质。由一条或多条多肽链组成的分子。可含有也可不含有辅基。

protein-coding gene 为蛋白质编码的基因。包含一个阅读框架的基因，其 mRNA 将被翻译成蛋白质。

provirus 前病毒。整合进宿主细胞基因组中的病毒基因组。

pseudogene 假基因。在顺序上与某一功能基因同源、但无功能的 DNA 片段。某一基因家族中的非功能性成员。

purifying selection(negative selection)纯洁化选择(负选择)。造成使某一等位基因从群体中去除这样的结果的选择体制。

purine 嘌呤。一类存在于核苷酸中的含氮碱基,由两个联在一起的环式结构组成,一个为五员环、另一个为六员环。DNA 和 RNA 中的嘌呤碱基为腺嘌呤和鸟嘌呤。

pyrimidine 嘧啶。存在于核苷酸中的一种含氮碱基类型,由一个六员环组成。DNA 中的嘧啶碱基是胞嘧啶和胸腺嘧啶。RNA 中的嘧啶碱基为胞嘧啶和尿嘧啶。

Q 谷氨酰胺。

quaternary structure 四级结构。具有两个或多个亚基的蛋白质分子中,两个或多个多肽链间相互作用的类型和方式,即代表了该蛋白质的四级结构。

R 精氨酸。

radical substitution 激进的替换。某一氨基酸被另一化学性质明显不同的氨基酸所替换。

random genetic drift(drift,genetic drift)随机遗传漂变(漂变、遗传漂变)。因偶然事件,例如配子的取样,而造成的等位基因频率随世代的波动。

rate of gene substitution 基因替换的速率。每单位时间每基因座位的基因替换数。

rate of mutation 突变率。每单位时间(通常以每世代时间)每基因座位或每核苷酸位点的突变数。

rate of nucleotide substitution 核苷酸替换速率。每单位时间每核苷酸位点的核苷酸替换数。

reading frame 阅读框架,读码框架。一个以起始密码子开头以终止密码子结束的为蛋白质编码基因中,密码子的直线顺序。

recessiveness 隐性。杂合子中某一等位基因不表现出来的现象。

recognition sequence 识别顺序。被某一限制性内切核酸酶识别的顺序。在很多情况下是一个短的回文顺序。

recombination 重组。染色体交叉后产生的结果,可看到顺反子中等位基因的新组合。

recombinator gene 重组子基因。一种给重组酶提供识别位点的调节基因。

reduction division(见 meiosis)减数分裂。

regional duplication 区域性重复。重复量小于全基因组的重复都称区域性重复。

regulatory gene 调节基因。一种非转录基因。有时用来指调节其他基因的表达的结构基因。

relative - rate test 相对速率测验。一种无需标度的测验。以检验不同谱系在其进化过程中核苷酸替换速率的恒定性,从而确定分子钟在不同谱系中是否以同样的速率运转。

repetitive DNA 重复 DNA。以许多拷贝存在于单倍基因组中的 DNA 序列。

replacement 替代、取代。在蛋白质水平上一次非同义替换所造成的结果。

replication 复制。DNA 在某一 DNA 模板上合成的过程。

replication slippage 复制滑脱。DNA 复制期间,某一 DNA 序列作为模板不止一次地连续起作用,从而在新合成的 DNA 上产生了一段串接重复的序列,这样的过程称复制滑脱。

replicative transposition(duplicative transposition)复制(型)转座。可转座因子的一个拷贝插入新的染色体位置,而该因子自身仍留在原初位置的现象。

replicator gene 复制子基因。一种指定 DNA 复制的起始和终止位点的调节基因。

replicon 复制子。一个染色体区域,它含有为 DNA 复制所必需的 DNA 序列,并且它作为一个单位而被复制。

reproductive barrier(reproductive isolation)生殖障碍(生殖隔离)。阻止群体间基因交换的几种生物学的、或环境的机制中的任何一种。

restriction endonuclease(restriction enzyme)限制性内切核酸酶(限制酶)。水解 DNA 的内部磷酸二酯键的酶。

restriction-fragment pattern 限制片段模式。某一 DNA 序列被限制性内切核酸酶消化后，它所产生的限制片段的数目和大小的情况。

restriction site 限制位点。限制性内切核酸酶水解（切割）DNA 的某一内部磷酸二酯键时的作用点。位置可能非常接近识别序列，也可能不接近识别位点。

restriction-site map 限制位点图谱。显示限制位点所处位置的 DNA 序列的图形。

retroelement 反录因子。具有产生反转录酶能力的 DNA 或 RNA 序列。

retrofection 反录感染（传染）。RNA 分子通过反录病毒粒子（RNA 被包在里面）而从一个细胞向另一个细胞（特别是种系细胞）转移，然后 RNA 被反转录并整合进宿主细胞，这样的过程称反杂感染。即通过反录病毒而转导。

retrogene（processed gene）反录基因（加工后基因）。一种有功能的反录序列，它产生的蛋白质与 mRNA 原来的基因所产生的蛋白质等同或接近等同。

retron 反录子。一种有反转录酶密码信息但缺乏转座能力的基因组序列。

retroposition 反录转座。一种由 RNA 中介的转座模式。

retroposon 反录座子，又译“反转录子”。一种既不构成病毒子粒子，又无两侧末端冗余序列的可转座反录因子。

retropseudogene（processed pseudogene）反录假基因（加工后假基因）。一种由 RNA 分子反转录而来，其后的 cDNA 整合进基因组而出现的假基因。判定反录假基因的标志有：缺乏内含子、多聚腺苷酸尾巴，两侧重复；表现出截尾等转录后修饰迹象；以及与该基因家族的有功能成员或非加工后的无功能成员都无连锁关系（无位置上的关联）。

retrosequence（retrotranscript，processed sequence）反录序列（反录本，加工后序列）一种由 RNA 反转录而来的基因组序列，但仅靠它自身则缺乏产生反转录酶的能力。反录基因和反录假基因都是反录序列。

retrotransposon 反录转座子。一种不构成病毒子粒子而两侧有末端冗余序列的可转座反录因子。

retrovirus 反录病毒。一类有反转录酶密码信息的小单链 RNA 病毒。

reverse transcriptase 反转录酶。催化逆向转录的酶。

reverse transcription 反转录。以 RNA 为模板而进行的单链 DNA 分子的合成。

ribonucleic acid（RNA）核糖核酸。由核苷酸连接而成的大分子多聚物，其中糖基为核糖。通常，RNA 呈单链。

ribosomal RNA（rRNA）核糖体 RNA。作为核糖体中的结构成份的 RNA 分子。

ribosome 核糖体。由 rRNA 和蛋白质组成的一种细胞内颗粒，为 mRNA 的翻译提供场所。

RNA（见 ribonucleic acid）

RNA-specifying gene 编码 RNA 的基因。可转录但产生的 RNA 不被翻译的基因。其 RNA 转录本是有功能者的基因。

rolling-circle replication 滚环复制。一种扩增模式，按此模式某一 DNA 序列的环状染色体外拷贝产生、并以连续方式复制。

root 根。在有根树中，所有被研究的分类单位的共同祖先。

root tree 有根树。一种指明祖先和后代物种，从而指出了进化途径的方向的系统树。

rRNA（见 ribosomal RNA）

S 丝氨酸。

secondary structure 二级结构。在蛋白质和核酸中，分别由氨基酸之间或核苷酸之间形成氢键而导致的结构。对蛋白质而言，是多肽链中的区域性结构（例如 α—螺旋、β—片状、转折）。对单链 DNA 或 RNA 来说，则为区域性双链结构（例如发卡）。

segregation（见 Mendelian segregation）分离。

segregation distortion（meiotic drive）分离偏斜（减数分裂驱动）。在杂合子的配子间出现的分离比偏离。

seqregator gene 分离子基因。一种在减数分裂和有丝分裂期间为分离机构提供染色体附着部位的调节基因。

selection(见 natural selection)选择。

selection coefficient 选择系数。某种基因型与群体中最适基因型相比,其适合度降低的定量测度。一种对选择不利性的测度。

selection intensity(stringency of selection)选择强度(选择的严峻程度)。群体中各种基因型间适合度值上的差异。

selective constraint(见 functional constraint)选择限制。

selfish DNA 自私 DNA。一种对其携带者(或宿主)可能无任何好处,而仅关心其自身传播的 DNA 片段。可转座因子被认为是自私 DNA。

self-splicing intron 自(我)拼接内含子。一种无需外来催化剂帮助而能从前 mRNA 中分裂出来的内含子。

sense codon 有义密码子。确定某种氨基酸的密码子。

sense strand 有义链。基因的非转录链,该链的 DNA 顺序与 RNA 转录本的等同。

sequence divergence(divergence)序列分岐(分岐)。两同源序列由于各自在其谱系中独立地积累遗传变化,而出现的相互间的差异。

sex-linkage 性连锁。位于性染色体上的基因所处的处境。该术语常限于指 x－染色体上的基因。

sibling species 姐妹种,同胞种。形态上不可区分但生殖上出现隔离的物种。

signal peptide 信号肽。在多肽合成后且在该蛋白质确定其正确的三级结构之前,而从多肽中分裂出来的领头肽。

silencing(见 nonfunctionalization)沉默(化)。

silent substitution 沉默替换。不改变其携带者的表型的替换。包括非基因 DNA 中的替换和同义替换。

simple transposon 简单转座子。不含插入序列的转座子。

SINE(Short Interspersed Element 的首字母缩略)短散在因子。多细胞真核生物的基因组中,长度短于 500 碱基对、拷贝数为 10^5 或更多的散在重复序列,都可称为 SINE(短散在因子)。

single-copy DNA(见 unique DNA)单拷贝 DNA。

sister taxa(neighboring taxa)姐妹分类单位(近邻分类单位)。一般用来指一群物种中相互在进化程度上最接近的物种对。在系统树中,即两个仅通过一个内部节点联系起来的分类单位。

somatic cell 体细胞。注定不会变成配子的细胞。

somatic mutation 体细胞突变。体细胞中发生的突变。

spacer DNA 间隔 DNA。处于两个基因之间的 DNA。可以被转录也可以不被转录。

speciation(cladogenesis)物种形成(分枝进化)。一个群体分裂成两个或多个生殖隔离的群体的过程。新物种通过此过程产生。

species 物种。分类中的一个基本范畴,对此有一些不同的定义:(1)一群真实的或潜在的可相互杂交的个体,它们与别的这样的类群生殖隔离(生物学物种概念);(2)一个与别的谱系在进化上分离的谱系(进化论物种概念);(3)一群彼此相似的生物,它们间的相似远胜于与群外生物的相似(分类学物种概念)。

species tree 物种树。表示一群物种间的进化关系的系统树。

splicing 拼接。在 RNA 走向成熟的加工中,去除内含子的过程。

splicing site or junction 拼接位点或拼接点。外显子和内含子间的界点。

split gene 断裂基因。含有内含子的基因。

standard nucleotide 标准核苷酸。即腺嘌呤核苷酸(A)、胞嘧啶核苷酸(C)、鸟嘌呤核苷酸(G)、胸腺嘧啶核苷酸(T)、或尿嘧啶核苷酸(U)。

sticky ends 粘性末端。从双螺旋 DNA 相对的两端伸出的 DNA 单链。通常是双链 DNA 被某种

限制酶错开式地切割而产生的。

stochastic process 随机过程。结果不能从初始条件知识而精确地预测到的过程即随机过程。不过，根据给定的初始条件，对该过程的每一可能出现的结果可得出一定的概率。

stop(见 termination codon)终止，休止符。

stringency of selection(见 selection intensity)选择的严峻程度。

strong bond 强键。与双键的核酸或片段(例如密码子—反密码子相互作用)有关的、存在于 C 和 G 间的三个氢键。强键能增加稳定性，提高熔解温度。

structural domain(见 module)结构域。

structural gene 结构基因。一段为蛋白质编码或确定一种 RNA 分子的 DNA 序列。另一些作者则仅把为蛋白质编码的基因当作结构基因。

subspecies 亚种。某一物种中的一个地理学上或形态学上有区别的群体。

substitution(见 gene substitution 和 nucleotide substitution)替换。

substitution matrix 替换矩阵。以矩阵的形式表示的替换模式，其中的元则表示某两个核苷酸间的相对替换速率。

substitution scheme(见 pattern of substitution)替换方式。

superfamily 超家族。相互间有一定程度分岐的一些基因的集合，把基因重复的一切产物都包括在内(对为蛋白质编码的基因来说，通常氨基酸水平上的相似性小于 50%)。

symbiosis 共生。两个或多个生物以互利关系共存。

symmetrical exon 对称外显子。位于两个同相位内含子之间的外显子。

synonymous substitution(silent substitution)同义替换(沉默替换)。核苷酸替换后的密码子确定的氨基酸与以前的相同，这样的替换称同义替换。

systematics 系统学。即分类学和系统发育学。

T (1)在 DNA 中表示胸腺嘧啶核苷。(2)在蛋白质中表示苏氨酸。

tandem duplication 串接重复。重复产物位于染色体上相互紧挨在一起处的重复形式。

taxon(复数 taxa)分类单位、分类群。指任何等级的分类学群(例如种、属、界)，生物个体可按分划标准而被归于各群之中。

taxonomy(classification)分类学(分类)。指将物种命名并归于分类学群中的原则和程序。

terminal node 末端节点。系统树中表示一个现存分类单位的图形。

termination codon(nonsense codon，stop)终止密码子(无意义密码子、休止符)。一种不存在与其对应的正常 tRNA 的密码子，其出现可终止翻译过程。普适遗传密码中的三个终止密码子为 UAG、UAA 和 UGA。

tertiary structure 三级结构。在蛋白质和核酸中，由其自身的折叠而导致的分子的三维结构即三级结构。

thermal stability 热稳定性。表示抗变性或抗熔解程度的一种性质。

Tm 双螺旋 DNA 熔解时的中间温度。

△*Tm* 同源双螺旋 DNA 与异源双螺旋 DNA 的中间熔解温度之间的差。

topology 拓扑学、拓扑图。系统树的分枝模式。

trans 反式。两个序列或基因排列在不同染色体上。

transcription 转录。在 DNA 模板上合成 RNA 分子。

transcription-initiation site(见 cap site)转录起始部位。

transcription-termination site 转录终止部位。RNA 转录终止所在的位点。与被多聚腺苷化了的 RNA 对照，该部位可能与多聚腺苷酸部位等同，也可能不同。

transduction 转导。宿主的遗传信息通过病毒从一个细胞向另一个细胞的转移。

transfer RNA(tRNA)转移 RNA。一种小分子核酸，带有一个反密码子，一个与特异氨基酸结合的位点，以及与核糖体和酶相互作用的识别位点。这里的酶起着将 tRNA 与特异氨基酸连结的作用。

transition 转换。嘌呤被嘌呤或嘧啶被嘧啶替换。

translation(decoding)翻译、转译(解码)。经 tRNA 中介而从 mRNA 分子的核苷酸顺序得到多肽的氨基酸顺序的过程,此过程是在与核糖体结合中发生的。

transposable element(mobile element)可转座因子(可移动因子)。一种能在某种生物的基因组中到处移动的序列。

transposition 转座。遗传物质从一个基因组位置向另一个位置的移动。

transposon 转座子。除了携带与转座功能有关的基因外还带有附加基因的可转座因子。

transversion 颠换。嘌呤被嘧啶替换,或相反情形的替换。

trisomy 三体。在别的染色体为两倍体的细胞中,某一染色体存在三个拷贝的现象。

tRNA(见 transfer RNA)。

true tree 真实树。代表了一群分类单位的真实进化史的系统树。

twofold degenerate site 二重简并位点。编码区中的某一核苷酸位点,若三种可能的核苷酸改变中有一种是同义的,而另两种是非同义的,则该位点即为二重简并位点。

U 尿嘧啶核苷。

unequal codon usage(biased codon usage)密码子的不等价应用(偏向性密码子应用)。在为蛋白质编码的基因里,某一密码子(家)族中的一个或多个密码子被不成比例地应用。

unequal crossing-over 不等价交换。两同源序列在不同染色体位置处交换,新产生的染色体出现某些区域的拷贝数不相同的现象。

unidentified reading frame(URF)未定阅读框架。一种其产物未知的 ORF。

unique DNA(single-copy DNA)单一 DNA(单拷贝 DNA)。在单倍基因组中仅以一个拷贝存在的 DNA 序列。

universal genetic code 普适(通用)遗传密码。左右绝大多数基因组的翻译的遗传密码。

unprocessed pseudogene 未加工的假基因。通过基因重复得到,而随后其中一个拷贝无功能化或沉默化,这样产生的假基因即未加工的假基因。

unrooted tree 无根树。一种既未确定根,也未指出进化途径的方向的进化树。

untranscribed sequence(见 flanking sequence)不转录序列。

upstream 上游。核酸上某一参考点的 5′ 方向(与转录方向相反)。

URF(见 unidentified reading frame)。

V 缬氨酸。

variable site or region 可变位点或可变区域。(1)在被比较的 DNA 序列中,由不同核苷酸占据的位点或区域。(2)严格地说,指 DNA 中能随时间而变的位点或区域。

variant repetition 变异的重复。自重复事件发生以来,基因重复的产物相互分岐,而出现略有差异的重复。

viability 生活力。一种适合度成份。某一给定基因型的个体从妊娠活至生殖年龄的概率。

virion 病毒子。一种病毒颗粒。

virus 病毒。一种微小的寄生生物,依赖宿主细胞复制其遗传物质和合成其蛋白质。其基因组可能是 RNA 也可能是 DNA,可以是单链的也可以是双链的。

W 色氨酸。

weak bond 弱键。与双链核酸(例如密码子—反密码子相互作用)有关的存在于 A 和 T、或 A 和 U 之间的二个氢键。弱键使热稳定性和熔解温度降低。

wild type 野生型。群体中某一基因座位上最常见的等位基因,只要这样的等位基因存在。

wobble pairing 摇摆配对。某些 tRNA 通过反密码子第一位与密码子第三位之间的非标准配对(例如,反密码子中的 U 与密码子中的 G 配对),而能识别一个以上的密码子,这样的配对称摇摆配对)。

x-linkage(见 sex-linkage)x 一连锁。

xenology 宿主学。作为某一水平基因转移事件的后果的序列相似性。

Y 酪氨酸。

zygote 合子。由单倍体配子细胞核融合产生的两倍体细胞。

主 题 索 引

D

E

H

J

K

L

M

N

P

Q

R

S

T

W

X

Y

Z

缩写词和种名索引

(陈建华译,张亚平、吴春花和李海鹏校)

Molecular Variation and Ecological Problems

分子变异与生态学问题

原作：T. Burke，W. Reiney 和 T. J. White

目　录

分子变异与生态学问题

1 引　　言

分子遗传学的新进展为生态学提供了许多新的并且具有很高价值的技术。就象这些技术似乎是解决生态遗传学和进化生态学问题的基础一样，它们也开始为其他生态学领域提供了有用的研究工具。现在有一种倾向，批评许多生态学家喜好和应用这些新技术是一种潮流，或可能是生态学中的最新"时尚"(Abrahamson 等 1989)，但我们认为这是生态学家们的积极响应，因为这种新的可能性实际上会引发一次关键性的技术飞跃。

这些技术直接的重要作用是有效地检测生物个体间的差别 DNA。组成所有 DNA 分子的四种核苷酸的显然很简单的线性排列与表现在以下几个方面巨大而复杂的异质性是很不相符合的，这些方面包括进化速率、差异很大的序列结构、突变速率、固定速率以及选择压力等(图 1)。许多生物体的一

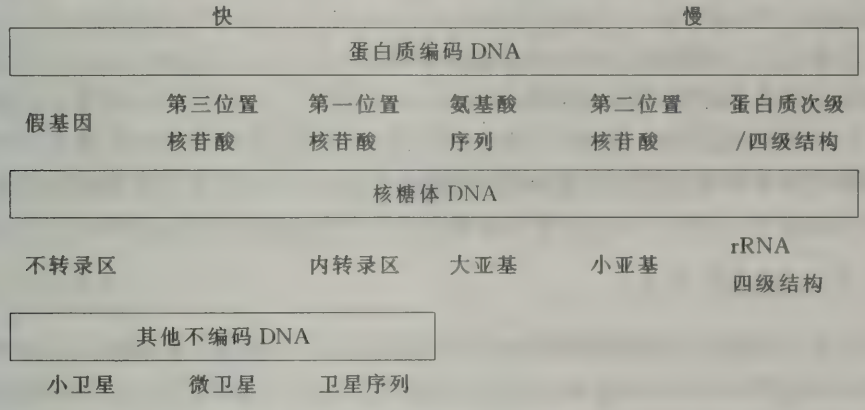


图 1　DNA 序列进化的相对速率

些基因组具有不同的遗传方式和序列进化率(见图 2)，对于它们的比较可以得到有价值的认识。另外，可用的各种分子技术本身提供了不同的敏感程度。这样就使从近缘的各种个体到在古代就已趋异的物种之间的每一种亲缘关系程度，最近已可能在目标 DNA 分级与适合作定量遗传差异(或相似)的分子方法间选择到一种结合。

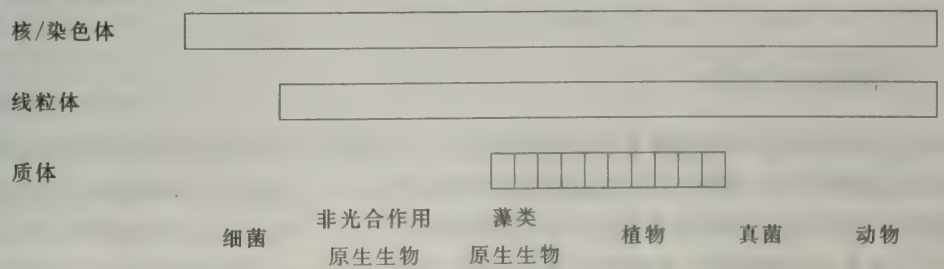


图 2　不同类群中可用作分析的基因组

长期以来，对表型变异与保持感兴趣的生态学家们，通过分离等位变异涉及表型的基因，能够直接分离和研究其遗传构成。这将与在商业上具重要价值的动、植物物种早已开展分子遗传研究一样，是一种自然的进步。

在这个综述中，我们主要集中于 DNA 变异的分析，而不是特异序列功能或特定位点选择过程的研究。我们通过有关生态学问题的实例来阐明其应用的范围。附表中，我们总结了迄今为止运用各种分子方法解决的生态学问题。我们顺序介绍其应用领域，并努力强调在每个领域中，现在看来最合适

的方法。由于大多数的技术可用于处理种种问题，为避免重复，首先我们来简单描述一下主要的新技术。

2 技术与术语

这一节论述了所用到的不同技术的一个概貌。与生态学家的需要特别有关的两本手册分别是由Heolzel(1992)和Hillis与Moritz(1990)编著的。本章所描述的已建立的主要方法在上述两本书上有详细的叙述和操作程序；其它的文献在需要的地方给出。

2.1 材料来源

实际上，任何细胞组织都是合适的DNA来源。在脊椎动物研究中，因为血液的获得是非破坏性的，所以是最常用的材料。DNA所需的量依所用的技术的不同而不同。例如，对于以DNA－DNA杂交为基础的方法，如多位点DNA指纹，每次样品操作需要5ugDNA(一般这个量足够以进行若干次操作)。材料从鸟类的血液(它的红细胞是具核的)中获得只需1－2ul，对于哺乳动物来讲，只有白细胞具核，则需要1ml血液。对于应用扩增技术的方法来讲，如聚合酶链式反应(PCR)，虽然由于存在潜在的污染问题，通常都避免这么小的取样，但是在理想条件下，开始于一单倍体基因组(haploid genome)的分析是可行的。

尽管多数实验室一直都是采用冷冻的方法保存用于提取DNA的组织，但已经证明各种化学混合剂(包括促溶剂，螯合剂，乙醇和浓盐溶液)能够在环境温度下保存适用于大多数分析方法的相对较长的DNA分子至少几周或几月(见Bruford等，1992)。例如，从酒精中保存6年之久的动物组织中提取出大量可被限制性核酸内切酶切断的DNA(Smith等1987)。但是另外的研究则报道了低的产量和快速降解(Seutin等，1990)。不同之处可能是由于商业酒精的污染(lto，1992)。

2.2 DNA－DNA杂交

杂交方法被用来(i)估计两个种基因组间总的相似性或(ii)在克隆之前，检测一个特殊的序列。它的简单方法是点印迹(dot blot)，即在与标记了的探针杂交前，将等分试样的DNA(甚至血液)点到滤膜上并固定。探针最常用的是放射性标记了的，但是非放射性标记的方法逐渐发展并普遍起来。通过各种合适的对照来比较其杂交强度可以测定目标样品中探针序列的存在。杂交的检测可以通过X射线放射自显影或直接在膜上探针的化学染色。当研究一特异性序列时，探针可以是一克隆了的DNA片段或一非常短的，典型的是20－30个碱基对，人工合成的寡核苷酸序列。一旦一个克隆被测序后，就可以设计寡核苷酸探针，并且原则上允许相对快速而方便地对序列变异数据的收集(如Gardes等1991)。

2.3 限制性片段分析

用点印迹法简单地检测相似序列经常是不够的，我们需要更详细的有关杂交的目标DNA的信息。直到最近，估测DNA序列变异的最简单方法是应用限制性片段(或限制性片段长度多态性，RFLP)来分析。首先，样品DNA被许多合用的限制性内切酶之一种切成一定的片段，然后依它们的长度将这些片段通过电泳分离。限制性片段可以通过电泳凝胶片的染色观察或如上所述，按Southern印迹法永久地转移到一个滤膜上经过杂交检测。(现在有些方法为了避免印迹这一步，而是把片段在凝胶上干燥后直接杂交凝胶，如Schafer等1988)。DNA片段的存在与否暗示着应用特定限制性内切酶来识别的特异性目标序列的变异。

2.4 DNA指纹分析

有两种不同级别的DNA指纹分析——多位点和单位点指纹。对于后者，通过与一克隆的小卫星探针的Southern印迹杂交，可以检测在一个单位点上等位DNA限制性片段的不同大小(Wong等

1987)。一个典型的小卫星 DNA 序列的长度大于 20,000 个碱基对,它包括一个短的 10－60 碱基对的非编码序列的重复拷贝。重复的次数有显著的差异,可以产生易于检测的不同长度的限制性片段,从而在一些小卫星位点得到了很高水平的多态性(杂合率达 100%)。小卫星位点是"可变数量的串联重复(VNTR)"位点的例子。

在多位点 DNA 指纹法中,用到了一个"多核心(poly-core)"探针,它可与部分重复单位(即"核心")自动杂交,这种重复单位在微卫星的很多分离的位点上是很普通的(Jeffrey 等,1985a,b)。可以用不同的核心探针来检测非依赖性小卫星带谱(independent minisatellite profiles)。多位点指纹法的一个主要优点是核心探针可用在广泛的生物体中(如 Burke 和 Bruford 1987, Dallas 1988,Taggart 和 Ferguson 1990,Carvalho 等 1991;见 Burke 等 1991a 综述)。有关多位点和单位点小卫星 DNA 指纹法的详细实验室操作程序见 Bruford 等(1992)的文章。

多位点指纹法的一个缺点是指纹谱复杂,难以比较,特别是在不同的凝胶上得到的谱。组成的 DNA 片段也不能归因于特异性位点。单位点指纹法通过允许基因型归因于常常是高度变异的位点避免了这些问题,但它也有缺点,就是在所研究的每一个物种中或至少在一个近缘种中都得找出一种标记系统(Hanotte 等,1991a,b,1992)。然而,现在可以使用一种相对有效的分离这种位点的方法(Armour 等 1990,Hanotte 等 1991a,b;Bruford 等(1992) 的操作程序)。

另外一种级别的 VNTR 位点包括了简单的序列,或微卫星多态性(Tautz1989)。微卫星(microsatellite)包含不同数量简单而短的串联重复,如(GT)n 或(CAC)n。小卫星和微卫星的区别是随意的,但实际的区别是微卫星总的长度小到允许应用 PCR 技术分析。所以微卫星的变异分析比小卫星容易得多(见 Rassmann 等 1991,Schiotterer 等(1991)的方法)。特异位点微卫星系统常在宽范围的近缘种中适用(Schlötterer 等,1991),在一定程度上比小卫星探针更有用。由于这里数据太少而不能在两个级别的两个位点进行典型变异的细致比较。高水平的多态性很可能在小卫星位点上得到,但微卫星的变异似乎适合作大量的应用。

2.5 DNA 放大

一个相对较新的方法应用于大多数 DNA 样品中——包括相当粗提的总 DNA——这个方法是聚合酶链式反应(PCR)(Mullis 和 Faloona, Saiki 等 1988)。这种技术可以产生足量的确定长度的 DNA,适合各种方法的进一步分析,如 RFLP、测序和与特异探针的杂交。在聚合酶链式反应中,应用一耐高温的 DNA 聚合酶循环复制相反的两条 DNA 链上两个引物部位之间的序列,这个反应需要两个合成的寡核苷酸引物的存在与 DNA 上的这些部位互补,每循环一次,产量成倍增加。这个过程的灵敏度使得它可以对极少量 DNA 进行分析,例如这些 DNA 可来自单个精子细胞(Amheim 等 1990),单根头发(Higuchi 等,1988),单个羽毛(Taberlet 和 Bouvet,1991),古代的骨胳(Hagelber,1989)或博物馆中的标本(Thomas 等,1989,1990)。由于所研究的序列只需要很少的完整拷贝 ,甚至固定的或包埋了的组织都适合来提取 DNA 作 PCR(Greer 等 1991),所以,很简单的野外样品保存和实验室提取的操作程序都会令人满意。

因为扩增的 PCR 产品有足够的量可供直接在电泳凝胶片上检测,所以 PCR 为杂交方法提供了另一个很好的选择。例如可以检测引物部位间的序列长度变异或检测在一个引物部位本身的变异。PCR 产物可以用作点印迹 DNA－DNA 杂交时的靶材料,也可用作限制性分析。它们和其它技术(下文)可以不需要广泛地测序就可以相对方便地收集有用的基于序列的标记数据。

2.6 DNA 测序

直到最近,得到一个序列之前,必须先从所研究的生物体中克隆一个序列。比如对整个基因的 DNA 多态性的研究 ,需要分别地从每个个体克隆基因 (如 Kreitman 1983)。一旦一特定位点至少在一个亲缘生物体中得到克隆并测得序列(由此可设计出寡核苷酸引物),PCR 的优点是通常可以绕过克隆这一步。对一些位点来说,可以鉴定两侧的序列段(stretches of flanking sequence),它们在一广阔范围的类群中是高度保守的,允许设计"通用"的引物(Kocher 等 1989,Hillis 和 Dixon 1991,Tabeer-

let 等 1991)。现在已经有了直接对 PCR 产物测序的很好的操作程序(如 Winship 1989,Innis 等 1990,Lee 1991)。一些工作者喜欢在测序前亚克隆 PCR 产物,但是这就意味着,特别是当多序列需要排除人工核苷酸替换的可能时,可能造成相当大的额外工作量。

2.7 变性梯度凝胶电泳

既便通过 PCR 和直接测序可以相对比较容易地获得序列数据,但是要积累起适合作核基因等位基因或线粒体单膜类型(haplotypes)比较的种群频率数据,是一个令人胆怯的任务。虽然除医学遗传学外,变性梯度凝胶电泳(DGGE)尚未被广泛应用,DGGE 及其相近的方法(如温度梯度凝胶电泳(TGGE)(Riesner 等 1989)和 PCR－单链构象多态性(SSCP)分析(Hayashi 1991))能够提供用来测序数十个样品的较低劳动强度的方法。Lessa(1992)和 Myers 等(1989)提供了 DGGE 方法详细的指导,但是,简单地说,这种方法需要为所研究的具有长 G-C 尾的区域准备一个引物。这种具有抗变性的 G-C 键的扩增产物可在含有脲的梯度凝胶上电泳。不同序列的产物在不同的脲浓度上部分地变性,对长度小于 500bp 的产物,小到只有 1 个碱基对替换的变化能够通过移动距离的不同而分辨出来。频率数据(包括杂合子的辨别)可直接由凝胶上得到。如果有要求的话,代表已检测了迁移等级的谱带上的样品可以被再放大并测序。在这种应用中,DGGE 对研究 DNA 片段在种群内或其它研究单位内的中等变异是最有用的(就是说,不是所有的个体都是独特的)。然而,这对从种群遗传或系统发育角度作初步探索可能有同等的价值,即在被选择的各种个体的 DNA 片段测序前,在种群或分类群内或之间去评估最初放大的 DNA 片段是否有一定程度的变异。

2.8 随机放大多态性 DNA

人们现在广泛地探索随机放大多态性 DNA(RAPD),部分是由于它显著的简单性。它依赖由被研究的生物体中提取的 DNA(有时来自分离的细胞器中),应用几组短的单链随机序列引物来测定变异(Williams 等 1990; Hardrys 等 1992)。经过琼脂糖电泳和溴化乙锭染色后,比较放大产物时,一些引物将产生少许分离的带,它们的一部分在种群内或种群间是多态的。当检测时,这些标记通常呈现显著的孟德尔遗传特性(Williams 等 1990, Arnold 等 1991)。在动植物中,这些标记可用来制作基因连锁图(Williams 等 1990)。

假设不了解被放大的区域,并且有关的书面出版物很少,对于每一种新的生物必须检验从不同个体中多次分别提取和放大(若有可能也包括遗传力)的产物模式的重复能力。应用得好,这种方法仅需要 DNA 分离、放大和琼脂糖凝胶电泳过程,就可提供丰富的多态性标记数据。避免了限制性酶和放射性标记的高消耗和复杂性。

3 生态学应用

在这一部分,我们讨论生态学家感兴趣的领域。所研究不同范畴的问题及用到的方法的实例列在附表中。我们试着尽可能地选择与生态学家有关系的实例并尽可能包括有代表性的分类群。在很多情况下,虽然一个新而有前途的技术建立了,但仍没有正式发表的应用实例,反映了这一领域中技术发展的步伐还不快。阅读附表时,应注意这种可能性。就象在附表中所列的一样,读者必须明白所选的一些实例并不代表最有效的方法。

3.1 性别鉴定

Fisher(1930)预言,在有性的物种中,一种亲本应该在它的每一性别的后代中有均等的总投入。Trivers 和 Willard(1973)争辩说,当亲本的投入影响到一个后代以后的生殖成功(reproductive success)并且在生殖成功中某一性别的变异较大时,那么一种亲本应该在这种性别有较大的投入。对于许多通常的生物体研究对象来说,由于难于在幼年期作性别鉴定(sexing juveniles),因此这些粗放的试验、相关的假设以及一般性的生活史的研究受到了障碍。有一个在形态上可区别性别的例子,东

部蓝色鸣鸟(*Sialia sialis*)的雄性亲本在它们的雌性后代中有较多的投入(Gowaty 和 Drage 1991)。有人认为,这个物种从出生到繁殖扩散之所以很低是因为嗜杀父(Phiopatric)的雄性鸟可以为了雌性而要与它的亲生父亲竞争,所以雌性后代将占雄性适合度较大的比例。

性别是由遗传决定的关键是至少在一种性别的本身具某些特异性 DNA。原则上这种情况只需一个位点的等位基因的差异,而实际上差异一般广泛得多,常常包括了具性别特异的染色体。在这些情况下细胞学的性别鉴定是可行的,但通常要求作大样品是不切实际的(见 Parker 等 1991)。用分子标记去鉴别特殊的性染色体的序列特征,已有了一些实例。例如,Griffiths 和 Holland(1990)使用一种减法克隆(subtraction cloning)技术分离了食鲱鱼鸥(*Larus argentatus*)的 W 染色体(雌性特异的)的特异性探针。遗憾的是,就象其它类似的探针一样,这个探针只对很少一些近缘种有应用价值。也偶尔地发现了对其它一些物种的伴性探针(如 Quinn 等 1990,Rabenold 等,1991)。

长期以来,鉴定和分析性别决定基因本身似乎避免了这个问题。最近分离了在 Y－染色体上的性别决定区(SRY)的基因,并认为它对人类来说是性别决定位点(Sinclair 等 1990, Gubbay 等 1990)。它能够作为一个探针用印迹分析来确定一系列哺乳动物的性别(Sinclair 等 1990)。虽然在其它脊椎动物分类群中,它好象可能具有一个相同功能的同源物,但首次在鸟类寻找这样一个同源基因并不成功(Griffiths 1991)。一旦得到了合适的序列数据,至少就可以设计和合成在大范围的哺乳动物物种中有用的寡核苷酸探针,或至少可以设计寡核苷酸引物应用于 PCR。

3.2 交配制度

进化生态学家对交配制度特别感兴趣,他们希望能够度量个体在自然条件下的生殖成功情况。或者,也需要测量互助个体间的亲缘关系。在实践中,父本和母本的检验是测量亲缘关系的一种特殊情况(如 Birkhead 等 1990)。为这个目的,应用分子方法——特别是多位点 DNA 指纹和随机 RFLP 分析——进行研究,可见 Burke(1989)的综述。最近,单位点指纹分析发展迅速,现在已成为可选用的方法(Burke 等 1991b)。

多位点分析已应用于一些完全在野外进行的繁殖系统研究上,现在几乎已经成为常规方法。迄今多数的研究集中在鸟类上,反映了鸟类作为行为生态学研究对象的普遍性,并且涉及到基本上是单配制度中确定是父性或是母性(见 Burke 等 1991b,Birkhead 和 Moller 1992)或者关注互助繁殖群成员之间的血统分布以及亲缘关系的程度(Burke 等 1989, Rabenold 等 1990,Jones 等 1991, Pacher 等 1991, Davies 等 1992)。

第一类研究范畴是关于交替交配对策(alternative mating strategies)的发生,例如,用排除分析(exclusion analyses)发现子代基因型与它们的推断亲本不一致,以此来检测额外配对(extra- pair) 交配或种内繁殖寄生现象(brood parasitism)。例如,在鸟类饲养场对斑纹鸣鸟(*Taeniopygia guttata*)的交配行为的研究导致期望额外配对的父本可以在自然种群中出现。在对一野外种群指纹的推断研究中,由行为指定(behaviourally-assigned)的双亲本是完全的,并且经过排除分析,确认额外配对的父本以较低的频率发生(82 个后代中有 2 个),而种内繁殖寄生现象则意外地较显著(80 个后代中有 9 个)(Birkhead 等 1990)。

例如,在一个目的是为说明家系的研究中,绝大多数岩鹨(*Prunella modularis*)一般为多雄受精繁殖,2 个或更多的雄性共同拥有一个雌性。这些雄性一般互相没有亲缘关系,但有一种显性的关系,可能其中一个或两个雄性一起去帮助雌性喂养雏鸟。虽然这些雄性对它们的后代没有明显的偏爱,但应用多位点指纹的分析表明如果它们存在某些家系关系,它们就会倾向于喂养同窝的雏鸟(Burke 等 1987)。如果观察到一只雄鸟在雌鸟已受精还未产卵前一段时期内排外性地去接近雌性,那么它就更倾向于帮助喂养,而且由雄性喂养的程度与排外性地接近雌鸟的程度显著相关。指纹分析的数据表明,用观察接近的途径是一个好的家系指示,所以这就意味着雄性岩鹨就是用它们接近雌鸟的程度来决定是否喂养雏鸟。随后在一系列的去除(removal)实验中,通过人工操作使雄性个体交配成功(通过 DNA 指纹来确认),以上观点被实验所证实(Davis 等 1992)。

就象在岩鹨的研究中的情况一样,在只有少数的雄性候选者的情况下,应用多位点指纹分析能

有效地指示家系，但是这经常需要对许多潜在的双亲作检验。在这种情况下，应用一个单位点指纹分析的系统是更有效的，它能在一系列高度多态位点上指示个体的基因型。Gibbs 等(1990)将多位点和单位点指纹分析结合起来，鉴别了 28%第 3 代红翅乌鸫幼鸟的几乎所有父本，这些幼鸟不是由拥有它们出生地的雄性所生的。这就说明了在一个雄性的领地上产生的后代数和它的实际生殖成功之间没有显著的相关，而且在它们自己领地上生育较大比例后代的这些雄性也进行了较多的额外交配的受精。Gibbs 及其合作者所用的单位点探针是一个鼠的组织相容性(histocompatibility)位点 cDNA，它的用途是通过与一包含限制性片段的小卫星的偶然杂交而发现的；但是没有发现它普遍能用于其它的鸟类(Gibbs，私人通信)。分别获得高多态性的小卫星或微卫星位点的特异性探针或引物的一般性的方法都是可行的(见前文)，而且这样一些位点结合使得多位点指纹法变得过时(Wong 等 1987，Hanotte 等 1991a)。其它的分子方法偶而也用到。Quinn 及其合作者们发现由雪雁中随机克隆的一系列 RFLP 有可观的变异，并且可用来进行这个种的家系分析(Quinn 和 White，1987，Quinn 等 1987)。Williams 和 Strobeck(1986)确认了果蝇 Y 染色体的核糖体基因的特异性 RFLPs，由此我们能够揭示在野外种群中雌性果蝇发生过多次交配。本领域更详细的论述已超出本文的范围，读者可参考一些已发表的综述(Burke 1989，Burke 等 1991b 和 Burke 等 1991a 中的其它论文)。

3.3 种群结构

种群结构是广义的话题，包括从社群(social groups) 内到种群间在各种水平上研究亲缘关系与遗传分化。在前面一节内，我们单独讨论了近缘家族关系的分析。

除等位酶不属于本评述的范围外，线粒体 DNA(mtDNA)是用于许多种群分化研究的大分子(Avise 等综述 1987)。例如，84 只座头鲸(humpback whale)的 mtDNA(由皮肤活体解剖得到)的 RFLP 分析证明，在亚种群之间，以及来自北大西洋和北太平洋的种群之间有明显的单模标本(haplotypes)的分离(Baker 等 1990)。单模标本的地理分布与先前报道的夏季取食地与冬季繁殖地之间的迁移模式有惊人的一致。研究者认为，这种遗传分离反映了"迁移目的的母系习性(materal traditions in migratory destination)"，揭示了要考虑到行为模式在种群结构分析中的重要性。在植物中，质粒 DNA 也提供了很多类似 mtDNA 同样的进展，并已在少数的种群结构研究中得到应用(如 Goff 和 Coleman 1988)。

从总基因组文库中分离的随机克隆已在种群水平上用于 RFLP 的研究。Mcdonald 和 Martinez (1990) 研究了从一块麦地得到的一种真菌病原体 *Mycosphaerella graminicola* 的种群结构。他们发现了高程度的遗传变异，包括一片叶子上不同的侵蚀斑(lesions)之间。他们的研究强调了在确认从不同地理位置取得的少量样品是否代表种群水平的变异之前，评估空间尺度变异的重要性。

也偶尔使用另外一些 DNA 标记类型，如核糖体 DNA(如 Learn 和 Schaal 1987)。尽管一般认为小卫星进化太快而不能用于种群水平的分析，Gilbert 等(1990)发现在讨论南加利福尼亚的海峡岛(channel islands)上矮小狐(dwarf foxes)小的分离种群的系统发育，以及分离种群间遗传变异的相关水平时，指纹模式能提供有益的信息。在大部分分离的岛上，他们发现在取样的狐之间没有指纹变异，一个指纹模式甚至在自交的实验室小鼠群中也未发现过。多位点指纹在鉴别自然种群内无性系成员方面也可有重要价值(Nybom 和 Schaal 1990，Turner 等 1990，Carvalho 等 1991，Brookfield 1992)。

PCR 使得在种群研究中收集实际的 DNA 序列是可行的，而且它也允许用收藏在博物馆中的标本作实验。Thomas 等(1990)利用这种有利条件，发现 78 年间 3 个啮(kangaroo rat)种群的种群结构没有明显的变化。

3.4 迁移和基因流

当然，迁移是种群结构的一个重要成分，所以，同样的方法学上的各种途径是有用的。在这个前提下，我们列举了一些实例来着重强调应用分子的方法有时可以推断迁移模式。

例如，在现代人的地理起源这个有争议的问题上，通过从多类人种和黑猩猩取样的 mtDNA 替位环区(displacement loop region)的序列比较，确认了人类 mtDNA 进化树的非洲起源(Vigilant 等

1991；例外见 Templeton 1992）。Hagelberg 等(1989)在研究人类的人口迁移时，扩展了 PCR 的用途。证明了扩增 5450 年以上的人骨的 mtDNA 的可行性。在一系列研究上，无疑地，对在形态学上无症状的骨头片段作了鉴定。

将 RFLP 分析与等位酶相结合，mtDNA 也被用在遗传学上研究和确认蝾螈两个种之间的历史上描绘的杂种地带(hybrid zone) 的变迁（Arntzen 和 Wallis 1991）。代替了有花纹的蝾螈(*Triturus marmoratus*)种群的有冠肉的蝾螈(*T. cristatus*)种群的特征是存在一种低的但可辨认的渐渗的 *marmoratus* 等位基因频率。

Culex pipiens 蚊通过非特异性过量的产生多种酯酶来抗有机磷酸酯杀虫剂。这些酯酶中有一个在 B 位点上用电泳可察觉到的扩增了的 B_2 等位基因。来自非洲、亚洲和北美洲样品的 B_2 酯酶结构基因序列的限制位点分析显示其高度的同源性。这就指出了 B_2 等位基因是由单突变发展而来，并通过对杀虫剂诱导选择的有利性，在世界范围种群之间快速扩散。

3.5 渐渗现象与杂交地带

适用于种群结构分析的技术也适于特殊情况的杂交地带。分析沿着杂交地带的基因渐渗，mtDNA 特别有价值(Harrison 1989 综述)。核糖体 DNA 也应用于某种程度的研究中(如 Baker 等 1989)。

Arnold 等(1991)用一种选择叶绿体基因的 PCR 和 RAPD 来研究路易斯安那鸢尾杂交种的形成。物种专有的 RAPD 标记的地理分布支持这样一个假设：*Iris fulva* 和 *I. hexagona* 通过基因流导致了定位的和扩散的渐渗。一个扩增的叶绿体基因片段(携带在雌性细胞质中)的 RFLP 分析，说明是由于传粉而不是种子散布导致了基因流。

这个领域提供了某些成功应用 DGGE 的首例。Lessa(1992)应用 DGGE 研究了一种地鼠的杂交地带。

为了了解在形态上还识别不了有杂种存在的地区，在有杂交潜势的白栎物种间基因交换的程度，Whittemore 和 Schaal(1991)观察了用来自矮牵牛叶绿体和大豆细胞核核糖体重复(ribosomal repeat)的克隆探针消化的基因组 DNA 限制片段印迹。他们发现，虽然叶绿体基因型有某些大尺度的地理变异，在同一地点一起生长的物种(包括常绿的和落叶的)却经常地具有相同的叶绿体基因型。一个种在不同的地点具有不同的叶绿体基因型，而细胞核核糖体的标记则显示了一种分布格局与所期望的物种界线(boundaries)是一致的。他们在叶绿体核 DNA 中观察到的地理变异和明显的基因流的极明显的悬殊差别的格局，强调了在依据由细胞器得来的数据解释种群或类群界线时要特别谨慎（Pamilo 和 Nei，1988）。

3.6 物种的鉴定

总 DNA 的提取物间杂交的程度已被广泛地作为一种方法用来估测物种间的亲缘关系。特别是在原核生物之间，当一些生物具 90%或以上的杂交值时，这种方法就用来把它们粗粗定为物种(Woese 1987)。然而，这种方法的分辨率和分类学范围经常是低的，需要作比较就要求有大量的 DNA。

物种特异性探针在研究宿主—媒介(vector)—寄生物系统中特别有用。Kukla 等(1987)使用包含有重复序列的放射活性标记的限制片段克隆，鉴定了采采蝇组织中锥虫的种和亚种。他们发展了一种应用于野外的方法，将分离的蝇腹部整个地放在尼龙膜上与标记的探针杂交，就能确认和鉴定肠锥虫。通过筛选一个基因文库，Harnett 等(1989)发现并测定了一寡核苷酸探针的序列，它能区别引起人类河盲症(river blindness)的一种线虫 *Onchocerca volvulus* 和形态上相似但不是病原体的另外一个 *Onchocerca* 种，它们携带在同样的蚋(blackfly)媒介上。

Persing 等(1990)应用包柔氏螺旋体(*Borrelia*)特异性引物和 PCR 去研究博物馆鹿标本上蜱中的症状 DNA 片段，表明 Lyme 疾病的临床发现至少 30 年前就已经在美国存在了。Lyme 病原体 *Borrelia burgdorferi* 另外一随机克隆的序列分析，能够帮助设计 PCR 引物，以此把北美类型与欧洲和亚洲隔

离类型分辨开(Rosa 等 1991)。

Goff 等(1988)应用探针研究 *Anaplasma marginale* 表面蛋白基因，检测了蜱与牛被传染的频率。由此可监测宿主携带状态、疾病传染方式和地方性兽病区蜱传染的流行。Rowan 和 Powers(1991)测定了由 22 个海洋动物宿主类群上取得的 131 个个体隔离群的单细胞藻类共生体的 rDNA 小亚基片段序列。系统发育上远缘的宿主上发现了近缘的藻类。这种折衷的和如此广的分布区对藻类共生体来说，是一种结合的群聚模式，这种模式与某些动物寄生物的分化历史有时同宿主一致的情况是十分不一样的(Page 1990)。

DNA 分析可以进行血液寄生节肢动物的寄主的分类。为更好地说明这种方法的潜力，Coulson 等(1990)应用单位点小卫星指纹分析蚊子的肠子内含物，得到了被这种蚊叮过的人的基因型。

3.7 系统学

生态学家之所以对系统学感兴趣，是考虑到去识别特殊的物种以及对物种间进化关系的理解。分子技术可广泛地应用于系统学的研究，这个领域已广泛地评述过了(Hillis 和 Moritz 1990，Hewitt 等 1991；表中的实例)。实际上，任何客观的、基于遗传的特性都具有可为分类学提供资料的意义，不同方法的价值将按所研究的群体内遗传多样性的大小而不同。尽管由 DNA－DNA 杂交提供的表型性的方法成功地应用着(特别是在鸟类中；Sibley 和 Ahlquist 1990)；DNA 序列的系统发育分析被看好，而且现在由于 PCR 的发明而使之变得更加可行。在这之前，只有核糖体 RNA 基因能测序到值得注意的程度(Woese 1987，Hillis 和 Dixon 1991)。

应用 PCR 的首次研究为测定 mtDNA 序列而研究了通用引物的特性(Kocher 等 1989)。例如，Richman 和 Price(1992)应用这样的引物，从 8 个 *Phylloscopus* 鸣禽同地种，得到了具有 910 个碱基对的线粒体细胞色素 b 序列，建立了这些群体的系统发育。支持这些物种的形态学、取食行为和生态学适应解释的一个比较分析中，系统发育被用来核实共同祖先的影响。

物种的分类与保护特别有关系，这一点已被一系列 mtDNA 变异的研究结果所强调过，在设想是分离的小地鼠(Learn 等 1982)和海滨麻雀(Avise 和 Nelson 1989)的物种或亚种间，在可能是杂种起源的红狼(Wayne 和 Jenks 1991) 以及被外源基因渐渗的濒危的美洲狮和灰狼种群(见 O'Brien 和 Mayr 1991)中，都见不到有 mtDNA 明显的分化。通过使用 PCR 对红狼标本的分析，排除了杂交是最近发生的可能性。分子系统学也用来揭示这样的事例——不同的分类群可被归并为一个种，特别是在形态上保守的谱系中(Daugherty 等 1990)。

3.8 群落多样性

多个 rDNA 小亚基序列的同时扩增提供了一个检测微生物自然多样性的有用方法。Giovannoni 等(1990)把从浮游细菌的自然种群得到的 16s rRNA 基因扩增部分的克隆测了序。经过培养，增加了可检测的多样性，一些以前未知的分类群既在系统发育上是分离的，又是微生物群落中在数量上重要的成员。

Ward 等(1990)简单地检测了一陆地温泉中蓝藻细菌丛中微生物物种的组成。用一保守的寡核苷酸引物，他们通过从环境样品中取得的 16s rRNA 合成了 DNA。经过克隆和测序，并与来自相似生境培养的有机体的 16s rRNA 序列相比较后发现，所检测的 8 个序列类型没有一个同这一地点以前已知的 14 种生物相一致。这清楚地表明，分子工具可以揭示出微生物群落中从前未检测出的多样性，而且新生物体的检测并不限于稀有类群。

4 结　论

DNA 水平变异的程度和类型，特别是在物种内，只是最近才为人所知。新技术特别是 PCR 的发明，使得检测这类变异变得容易起来。一般来讲，以 PCR 为基础的方法已经成为并且将来仍是我们所选择的研究方法。

出于对一些有利因素的考虑，mtDNA 已经成为在 DNA(限制片段)水平上研究不同种群的通用信息源(见 Hillis 和 Moritz 1990)。这延伸了由 PCR 获得序列数据的情况，原因可能是由于 mtDNA 分子具有比核 DNA 序列更经常的核苷酸置换率；对 mtDNA 一些区域的了解和高度不同的进化速率；"通用"引物的可用性(Kocher 等 1989)以及线粒体基因组的单倍性。最后一点避免了在双倍体核基因组中潜在的问题，也避免了测定单模标本的序列(Kreitman 1991)。通常核标记也是可行的(Pamilo 和 Nei 1988)，虽然在这里经常用到"等位酶"，而 DNA 标记在确定变异程度时能提供更大的灵活性。由于物种内典型的核苷酸多态性水平是 1%或更小(Kreitman 1991)，对生态学家来说，样品的大量序列分析一般不是最适合的方法。于是，我们转向更简捷的方法，例如，应用 DGGE，寡核苷酸探针，限制性酶切片段，或者对已测序的多态区中的 PCR 产物的简单的大小测度。

虽然这些技术变的易于掌握，但它们的成功应用仍需要很多的脑力劳动，而且生态学家要注意，不要陷入只把实验室工作看作简单技术活动的错误之中。最后，分子方法在生态学过程研究中的应用，将是由不辞辛劳，互相了解对方优缺点的科学家们将野外和实验室工作巧妙结合的结果。

参考文献

Abrahamson, W. G. , Whitham, T. G. & Price, P. W. (1989) . Fads inccology. BioScience, 39, 321－325.

Armour, J. A. L. , Povey, S. , Jeremiah, S. & Jeffreys, A. J. (1990). Systematic cloning of human minisatellites from ordered array charomid libraries. Genomics, 8, 501－502.

Arnheim, N. , Li , H. & Cui, X. (1990). PCR analysis of DNA sequences in single cells : single sperm gene mapping and genetic disease diagnosis. Genomics, 8, 415－419.

Arnold, M. L. , Buckner, C. M. & Robinson, J. J. (1991) . Pollen mediated introgression and hybrid speciation in Louisiana irises. Proceedings of the National Academy of Sciences, USA, 88, 1398－1402

Arntzen, J. W. & Wallis, G. P. (1991). Restricted gene flow in a moving hybrid zone of the newts Triturus cristatus and T. marmoratus in western France. Evolution, 45, 805－826.

Avise, J. C & Nelson, W. S(1989). Molecular relationships of the extinct dusky scaside sparrow. Science, 243, 646－648.

Avise, J. C. , Arnold, R. M. , Ball, E. , Bermingham, T. , Lamb, G. E. , Neigel, C. A. , Reeb, C. A. & Saunders, N. C. (1987) . Intraspecific phylogeography: the mitochondrial bridge betwwen population genetics and systematics. Annual Review of Ecology and Systematics, 18, 489－522

Baker, C. S. , Palumbi, S. R, , Lambertsen, R. H. , Weinrich, M. T. , Calambokidis, J. & O'Brien, S. J. (1990) . Influence of seasonal migration on geographical distribution of mitochondrial DNA haplotypes in humpback whales. Nature, 344, 238－240.

Baker, R. J. , Davis , S. K. , Bradley, R. D, . Hamilton, M. J. & Van Den Bussche, R. A. (1989) . Ribosomal-DNA, mitochondrial-DNA, chromosomal, and allozymic studies on a contact zone in the pocket gopher, Geomys. Evolution, 43, 63－75.

Balazs, I. , Neuweiler, J. , Gunn, P. , Kidd, K. K. , Kuhl, J. & Mingjun, L. (1992). Human population genetic studies using hypervariable loci. Genetics, 131, 191－198.

Birkhead, T. R. & Moller, A. P. (1992) . Sperm Competition in Birds. Academic Press, London, UK.

Birkhead, T. R. , Burke , T. , Zann, R. , Hunter, F. M. & Krupa, A. P. (1990). Extra-pair paternity and intraspecific brood parasitism in wild zebra finches, Taeniopygia guttata, revealed by DNA fingerprinting. Behavioural Ecology and Sociobiology, 27, 315－324.

Brookfield, J. F. Y. (1992) . DNA fingerprinting in clonal organisms. Molecular Ecology, 1, 21－26.

Bruford , M. W. , Hanotte, O. , Brookfield, J. F. Y. & Burke, T. (1992) . Single locus and multilocus DNA fingerprinting. Molecular Genetic Analysis of Populations: A Practical Approach (Ed. by A. R. Hocelzel), pp. 225－269. IRL Press, Oxford, UK.

Bruns, T. D. , White, T. J. & Taylor, J. W. (1991). Fungal molecular systematics. Annual Review of Ecology and Systematics, 22, 525－564

Burke, T. (1989). DNA fingerprinting and other methods for the study of mating success. Trends in Ecology and Evolution, 4, 139－144.

Burke, T. & Bruford, M. W. (1987). DNA fingerprinting in birds. Nature, 327, 149－152.

Burke, T. , Davies, N. B. , Bruford, M. W. & Hatchwell, B. J. (1989) . Parental care and mating behaviour of polyandrous dunnocks Prunella modularis related to paternity by DNA fingerprinting. Nature, 338, 249－251.

Burke, T. , Dolf, G. , Jeffreys, A. J. & Wolff, R. (Eds) (1991a) . DNA Fingerprinting: Approaches and Applications. Birkhäuser, Basel, Switzerland.

Burke, T. , Hanotte, O. , Bruford, M. W. & Cairns, E. (1991b) . Multilocus and single locus minisatellite analysis in population biological studies. DNA Fingerprinting: Approaches and Applications (Ed. by T. Burke, G. Dolf, A. J. Jeffreys & R. Wolff), pp. 154－

168. Birkhäuser, Basel, Switzerland.

Carvalho, G. R., Maclean, N., Wratton, S. D., Carter, R. E. & Thurston, J. P. (1991). Differentiation of aphid clones using DNA fingerprinting from individual aphids. Proceedings of the Royal Society of London, B. 243, 109—114.

Coulson, R. M. R., Curtis, C. F., Ready, P. D., Hill, N. & Smith, D. (1990). Amplification and analysis of human DNA present in mosquito bloodmeals. Medical and Veterinary Entomology, 4, 357—366.

Dallas, J. F. (1988). Detection of DNA fingerprints of cultivated rice by hybridization with a human minisatellite probe. Proceedings of the National Academy of Sciences, USA, 85, 6831—6835.

Daugherty, C. H., Cree, A., Hay, J. M. & Thompson, A. M. B. (1990). Neglected taxonomy and continuing extinctions of tuatara (Sphenodon). Nature, 347,177—179.

Davies ,N B., Hatachwell, B. J., Robson, T. & Burke, T. (1992). Paternity and parental effort in dunnocks Prunella modularis:how good are male chick-feeding rules? Animal Behaviour, 43, 729—745.

Fisher, R. A. (1930). The Genetical Theory of Natural Selection. Clarendon Press, Oxford, UK.

Gardes, M., White, T. J., Fortin, J. A., Bruns, T. D. & Taylor, J. W. (1991). Identification of indigenous and introduced symbiotic fungi in ectomycorrhizae by amplification of nuclear and mitochondrial ribosomal DNA. Canadian Journal of Botany, 69, 180—190.

Gibbs, H. L., Weatherhead, P. J., Boag, P. T., White, B. N. ,Tabak, L. M. & Hoysak, D. J. (1990). Realized reproductive success of polygynous red-winged blackbirds revealed by DNA markers. Science,250, 1394—1397.

Gilbert, D. A., Lehman, N., O'Brien, S. J. & Wayne, R. K. (1990). Genetic fingerprinting reflects population differentiation in the California Channel Island fox. Nature,344,764—767.

Giovannoni,S. J. ,Britschgi,T. B. ,Moyer,C. L. & Field,K. G. (1990). Genetic diversity in Sargasso Sea bacterioplankton. Nature, 345, 60—63.

Goff, L. J. & Coleman, A. W. (1988). The use of plastid DNA restriction endonuclease patterns in delineating red algal species and populations. Journal of Phycology, 24,357—368.

Goff, W., Barbet, A., Stiller, D., Palmer, G., Knowles, D., Kocan, K., Gorham,J. & Meguire, T. (1988). Detection of Anaplasma marginale-infected tick vectors by using a cloned DNA probe. Proceedings of the National Academy of Sciences, USA, 85, 919—923.

Gowaty , P. A. & Droge, D. L. (1991). Sex ratio confiliet and the evolution of sex-biased provisioning in birds. Acta XX Congressus Internationalis Ornithologici, II,932—945.

Greer, C. E., Lund, J. K. & Manos, M. M. (1991). PCR amplification of paraffinembedded tissues: recommendations on fixatives for long term storage and prospective studies. PCR Methods and Applications, I, 46—50.

Griffiths, R. (1991). The isolation of conserved DNA sequences related to the human sexdetermining region Y gene from the lesser black-backed gull(Larus fuscus). Proceedings of the Royal Society of London, B, 244,123—128.

Griffiths, R. & Holland, P. (1990). A novel avian W chromosome DNA repeat sequenee in the lesser black-backed gull(Larus fuscus). Chromosoma, 99, 243—250.

Gubbay, J., Collignon, J., Koopman, P., Capel, B., Economou,A., Munsterberg, A., Vivian, N., Goodfellow, P. & Lovell-Badge, R> (1990). A gene mapping to the sexdetermining regions of the mouse Y chromosome is a member of a novel family of embryonically expressed genes. Nature, 346, 245—250.

Hadrys, H., Ballick, M. & Schierwater, B. (1992). Applications of random amplified polymorphic DNA in molecular ecology. Molecular Ecology , I, 55—63.

Hagelberg, E., Sykes, B. & Hedges, R. (1989). Ancient bone Dna amplified. Nature, 342—485.

Hanotte,O. ,Burke,T. ,Armour,J. A. L. & Jeffreys,A. J. (1991b). Cloning,characterization and evolution of Indian peafowl Pavo cristatus minisatellite loci. DNA Fingerprinting:Approaches and Applications (Ed. by T. Burke,G. Dolf,A. J. Jeffreys & R. Wolff),pp. 193—216. Birkhäuser,Basel,Switzerland.

Hanotte, O., Burke, T., Armour, J. A. L. & Jeffreys, A. J. (1991a). Hypervariable minisatellite DNA sequences in the Indian peafowl Pavo cristatus. Genomics, 9, 587—597.

Hanotte, O., Cairns, E., Robson, T., Double, M. & Burke, T. (1992). Cross-species hybridization of a single locus minisatellite probe in passerine birds. Molecular Ecology, I, 127—130.

Harnett, W., Chambers, A. E., Renz, A. & Parkhouse, R. M. E. (1989). An oligonuclcotide specific for Onchocerca volvulus. Molecular and Biochemical Parasitology, 28, 77—84.

Harrison, R. G. (1989). Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. Trends in Ecology and Evolution, 4, 6—11.

Hayashi, K. (1991). PCR-SSCP: A simple and sensitive method for detection of mutations in the genomic DNA. PCR Methods and Applications, 1, 34—38.

Helm-Bychowski, K. M. & Wilson, A. C. (1986). Rates of nuclear DNA evolution in pheasant-like birds: evidence from restriction

maps. Proceedings of the National Academy of Sciences, USA, 83, 688—692.

Hewitt, G. M., Johnston, A. W. B. & Young, J. P. W. (Eds) (1991). Molecular Techniques in Taxonomy. Springer Verlag, Berlin, Germany.

Higuchi, R., von Beroldingen, C. H., Sensabaugh, G. F. & Erlich, H. A. (1988). DNA typing from single hairs. Nature, 332, 543—546.

Hillis, D. M. & Dixon, M. T. (1991). Ribosomal DNA: molecular evolution and phylogenetic inference. Quarterly Review of Biology, 66, 411—453.

Hillis, D. M. & Moritz, C. (Eds)(1990). Molecular Systematics. Sinauer, Sunderland, MA, USA.

Hoelzel, A. R. (Ed.) (1992). Molecular Genetic Analysis of Populations: A Practical Approach. IRL Press, Oxford, UK.

Innis, M. A., Gelfand, D. H., Sninsky, J. J. & White, T. J. (Eds) (1990). PCR Protocols: A Guide to Methods and Applications. Academic Press, New York, NY, USA.

Ito K. (1992). Nearly complete loss of nucleic acids by commercially available highly purified ethanol. Biotechniques 12, 69—70.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985a). Hypervariable 'minisatellite' regions in human DNA. Nature, 314, 67—73.

Jeffreys, A. J., Wilson, V. & Thein, S. L. (1985b). Individual-specific 'fingerprints' of human DNA. Nature, 316, 76—79.

Jones, C. S., Lessells, C. M. & Krebs, J. R. (1991). Helpers-at-the-nest in European beecaters (Merops apiaster): a genetic analysis, DNA Fingerprinting: Approaches and Applications (Ed. by T. Burke, G. Dolf, A. J. Jeffreys & R. Wolff), pp. 169—192. Birkhäuser, Basel, Switzerland.

Kocher, T. D., Thomas, W. K., Meyer, A., Edwards, S. V., Pääbo, S., Villablanca, F. X. & Wilson, A. C. (1989). Dynamics of mitochondrial evolution in animals: amplification and sequencing with conserved primers. Proceedings of the National Academy of Sciences, USA, 86, 6196—6200.

Kreitman, M. (1983). Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature, 304, 412—417.

Kreitman, M. (1991). Variation at the DNA level: something for everyone. Molecular Techniques in Taxonomy (Ed. by G. M. Hewitt, A. W. B. Johnston & J. P. W. Young), pp. 15—32. Springer Verlag, Berlin, Germany.

Kukla, B. A., Majiwa, P. A. O., Young, J. R., Moloo, S. K. & ole-Moiyoi, O. (1987). Uses of species-specific DNA probes for detection and identification of trypanosome infection in tsetse flies. Parasitology, 95, 1—16.

Laerm, J., Avise, J. C., Patton, J. C. & Lansman, R. A. (1982). Genetic determination of the status of an endangered specics of pocket gopher in Georgia. Journal of Wildlife Management, 46, 513—518.

Learn, Jr. G. H. & Schaal, B. A. (1987). Population subdivision for ribosomal repeat variants in Clematis fremontii, Evolution, 41, 433—438.

Lee, J. S. (1991). Alternative dideoxy sequencing of double-stranded DNA by cyclic reactions using Taq polymerase. DNA and Cell Biology, IO, 67—73.

Lessa, E. P. (1992). Analysis of DNA sequence variation at the population level by PCR and denaturing gradient gel electrophoresis. Methods in Enzymology (in press).

Lucchini, G. M. & Altwegg, M. (1992). Ribosomal-RNA gene restriction patterns as taxonomic tools for the genus Aeromonas. International Journal of Systematic Bacteriology, 42, 384—389.

McDonald, B. A. & Martinez, J. P. (1990). DNA restriction fragment leugth polymorphisms among Mycosphaerella graminicola isolates collected from a single wheat field. Phytopathology, 80, 1368—1373.

Moritz, C., Dowling, T. E. & Brown, W. M. (1987). Evolution of animal mitochondrial DNA: relevance for population biology and systematics. Annual Review of Ecology and Systematics, 18, 269—292.

Mullis, K. B. & Faloona, F. A. (1987). Specific synthesis of DNA in vitro via a polymerase catalysed chain reaction. Methods in Enzymology, 55, 335—350.

Myers, R. M., Sheffield, V. C. & Cox, D. R. (1989). Mutation detection by PCR, GC-clamps and denaturing gradient gel electrophoresis. PCR Technology: Principles and Applications for DNA Amplification (Ed. by H. A. Erlich), pp. 71—88. Stockton Press, New York, NY, USA.

Nybom, H. & Schaal, B. A. (1990). DNA 'fingerprints' reveal genotypic distributions in natural populations of blackberries and raspberries (Rubus, Rosaceae) American Jouranl of Botany, 77, 883—888.

O'Brien, S. J. & Mayr, E. (1991). Bureaucratic mischief: recognizing endangered species and subspecies. Science, 251, 1187—1188.

Olson, R. R., Runstadler, J. A. & Kocher, T. D. (1991). Whose larvae? Nature, 351.

Packer, C., Gilbert, D. A., Pusey, A. E. & O'Brien, S. J. (1991). A molecular genetic analysis of kinship and cooperation in African lions. Nature, 352, 562—565.

Page, R. D. M. (1990). Temporal congruence and cladistic analysis of biogeography and cospeciation. Systematic Zoology, 39, 205—226.

Palmer, J. D. (1987). Chloroplast DNA evolution and biosystematic uses of chloroplast DNA variation. American Naturalist, 140, S6—S29.

Pamilo, P. & Nei, M. (1988). Relationships between gene trees and species trees. Molecular Biology and Evolution, 5, 568—583.

Parker, J. S., Birkhead, T. R., Joshua, S. K., Taylor, S. & Clark, M. S. (1991). Sex ratio in a population of guillemots Uria aalge determined by chromosome analysis. Ibis, 133, 423—424.

Persing, D. H., Telford, ▮ S. R., Rys, P. N., Dodge, D. E., White, T. J., Malawista, S. E. & Spielman, A. (1990). Detection of Borrelia burgdorferi DNA in museum specimens of Ixodes dammini ticks. Science, 249, 1420—1423.

Quinn, T. W. & White, B. N. (1987). Identification of restriction-fragment length polymorphisms in genomic DNA of the lesser snow goose(Anser caerulescens). Molecular Biology and Evolution, 4, 126—143.

Quinn, T. W., Cooke, F. & White, B. N. (1990). Molecular sexing of geese using a cloned Z chromosomal sequence with homology to the W chromosome. Auk, 107, 199—202

Quinn, T. W., Quinn, J. S., Cooke, F. & White, B. N. (1987). DNAmarker analysis detects multiple maternity and paternity in single broods of the lesser snow goose(Anser caerulescens). Nature, 326, 392—394.

Rabenold, P. P., Piper, W. H., Decker, M. D. & Minchella, D. J. (1991). Polymorphic minisatellite amplified on avian W chromosome. Genome, 34, 489—493.

Rabenold, P. P., Rabenold, K. N., Piper, W. H., Haydock, J. & Zack, S. W. (1990). Shared paternity revealed by genetic analysis in cooperatively breeding tropical wrens. Nature, 348, 538—540.

Rassmann, K., Schlütterer, C. & Tautz, D. (1991). Isolation of simple sequence loci for use in polymerase chain reaction-based DNA fingerprinting. Electrophoresis, 12, 113—118.

Raymond, M., Callaghan, A., Fort, P. & Pasteur, N. (1991). Worldwide migration of amplified insecticide resistance genes in mosquitoes. Nature, 350, 151—153.

Richman, A. D. & Price, T. (1992). Evolution of ecological differences in the Old World leaf warblers : roles of history and adaptation. Nature, 355, 817—821.

Riesner, D., Steger, R., Zimmat, R., Owens, R. A., Wagenhofer, M., Hillen, W., Vollbach, S. & Henco, K. (1989). Temperature- gradient gel electrophoresis of nucleic acids: Analysis of conformational transitions, sequence variations and protein-nucleic acid interactions. Electrophoresis, IO, 377—389.

Rogstad, S. H. Surveying plant genomes for VNTR loci. Molecular Evolution: Producing the Biochemical Data(Ed. by E. A. Zimmer, T. J. White, R. L. Cann, & A. C. Wilson). Academic Press, San Diego, CA, USA (in press).

Rosa, P. A., Hogan, D. & Schwan, T. G. (1991). Polymerase chain reaction analyses identify two distinct classes of Borrelia burgdorferi. Journal of Clinical Microbiology, 29, 524—532.

Rowan, R. & Powers, D. (1991). A molecular genetic classification of zooxanthellae and the evolution of animal algal symbioses. Science, 251, 1348—1351.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J. Higuchi, R., Horn, G. T., Mullis, K. B. & Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science, 239, 487—491.

Schäfer, R., Zischler, H., Birsner, U., Becker, A. & Epplen, J. T. (1988). Optimized oligonucleotide probes for DNA fingerprinting. Electrophoresis, 9, 369—374.

Schlötterer, C., Amos, B. & Tautz, D. (1991). Conservation of polymorphic simple sequcnce date in ceacean species. Nature, 354, 63—65.

Seutin, G., White, B. N. & Boag, P. T. (1990). Preservation of avian blood and tissues for DNA analyses. Canadian Journal of Zoology. 69, 82—90.

Sibley, C. J. & Ahlquist, J. E. (1990). Phylogeny and Classification of Birds: A study in Molecular Evolution. Yale University Press, New Haven, CT, USA.

Sinclair, A. H., Berta, P., Palmer, M. S., Hawkins, J. R., Griffiths, B., Smith, M. J., Foster, J. W., Frischauf, A. M., Lovell-Badge, R., Goodfellow, P. N. (1990). A gene from the human sex-determining regions encodes a protein with homology to a conserved DNA-binding motif. Nature, 346, 240—244.

Smith, L. J., Braylan, R. C., Nutkis, J. E., Edmundsen, K. B., Downing, J. R. & Wakeland, E. K. (1987). Extraction of cellular DNA from human cells and tissues fixed in ethanol. Analytical Biochemistry, 160, 135—138.

Soltis, P. S., Soltis, D. E. & Doyle, J. J. (1992). Plant Molecular Systematics. Chapman & Hall, London, UK.

Taberlet, P. & Bouvet, J. (1991). A single plucked feather as a source of DNA for bird genctic studies. Auk, 108, 959—960.

Taberlet, P., Gielly, L., Pautou, G. & Bouvet, J. (1991). Universal primers for amplification of three non-coding regions of chloroplast DNA. Plant Molecular Biology, 17, 1105—1109.

Taggart, J. B. & Ferguson, A. (1990). Minisatellite DNA fingerprints of salmonid fishes. Animal Genetics, 21, 377—389.

Tautz, D. (1989). Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research, 17,

6463—6471.

Templeton, A. R. (1992). Human origins and analysis of mitochondrial DNA sequences. Science, 255,737.

Thomas, R. H. & Hunt, J. A. (1991). The molecular evolution of the alcohol dehydrogenase locus and the phylogeny of Hawaiian Drosophila. Molecular Biology and Evolution, 8, 687—702.

Thomas, R. H., Schaffner, W., Wilson, A. C. & pääbo, S. V. (1989). DNA phylogeny of the extinct marsupial wolf. Nature, 340, 465—467.

Thomas, W. K., Pääbo, S., Villablanca, F. X. & Wilson, A. C. (1990). Spatial and temporal continuity of kangaroo rat populations shown by sequencing mitochondrial DNA from museum specimens. Journal of Molecular Evolution, 31, 101—112.

Trivers, R. L. & Willard, D. E. (1973). Natural selection of parental ability to vary the sex ratio of offspring. Science, 179, 90—92.

Turner, B. J., Elder, J. F. Jr, Laughlin, T. F. & Davis, W. P. (1990) Genetic variation in clonal vertebrates detected by simple-sequence DNAfingerprinting. Proceedings of the National Academy of Sciences. USA, 87, 5653—5657.

Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. (1991). African populations and the evolution of human mitochondrial DNA. Science, 253, 1503—1507.

Ward, D. M., Weller, R. & Bateson, M. M. (1990). 16S rRNA sequences reveal numerous uncultured mieroorganism in a natural community. Nature, 345, 63—65.

Wayne, R. k. & Jenks, S. M. (1991). Mitochondrial DNA analysis implying extensive hybridization of the endangered red wolf Canis rufus. Nature, 351, 565—568.

Welsh, J., Pretzman, C., Postic, D., Saint Girons, I., Baranton, G. & McClelland, M. (1992). Genomic fingerprinting by arbitrarily primed polymerase chain reaction resolves Borrelia burgdorferi into three distinet phylogenetic groups. International Journal of Systematic Bacteriology, 42, 370—377.

Whittemore, A. T. & Schaal, B. A. (1991). Interspecific gene flow in sympatric oaks. Proceedings of the National Academy of Sciences, USA, 88, 2540—2544.

Williams, J. G. K., Kubelik, A. R., Livak, K. J., Rafalski, J. A. & Tingey, S. V. (1990). DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Research, 18, 6531—6535.

Williams, S. M. & Strobeck, C. (1986). Measuring the multiple insemination frequency of Drosophila in nature: use of a Y-linked molecular marker. Evolution, 40, 440—442.

Winship, P. R. (1989). An improved method for directly sequencing PCR amplified material using dimethyl sulphoxide. Nucleic Acids Research, 17, 1266.

Woese, C. (1987). Bacterial evolution. Microbiological Reviews, 51, 221—271.

Wong, Z., Wilson, V., Patel, I., Povey, S., Jeffreys, A. J. (1987). Characterization of highly variable minisatellites cloned from human DNA. Annals of Human Genetics, 51, 269—288.

（魏伟译自 T. Burke, W. E. Reiney and T. J. White, 1992, Molecular Variation and Ecological Problems, in R. J. Berry, T. J. Crawford & G. M. Hewitt (eds.)《Genes in Ecology》229—254, Oxford: Blackwell Scientific Publications 钱迎倩 校）

2000年系统学议程:制订生物圈计划

——全世界物种的发现、描述和分类的全球计划技术报告

美国植物分类学家学会、系统生物学家学会、Willi Hennig 学会和系统学标本馆联合会
组成的联合体

Systematics Agenda 2000: Charting the Biosphere

A Global Initiative to Discovery, Describe and Classify the World's Species

the American Society of Plant Taxonomists, the Society of Systematic Biologists and
the Willi Hennig Society in cooperation with the Association of Systematics Collections, 1994

目　录

2000年系统学议程:制订生物圈计划

——全世界物种的发现、描述和分类的全球计划技术报告

前　言

遐想你被置身于一个陌生但又美丽的星球。四周景色宜人。漫步沙滩,翻滚的波浪拍打着海岸,岸上绿草悠悠,溪流潺潺,茂密的森林衬托着远处白皑皑的雪山。你只消一瞥便能领悟到这是个充满生命的世界——种类如此之多令人感到茫然,乍看约或令人发颤。

试想会有多少东西能从这个新的星球了解到,而这些知识会有多么的重要。有些植物可能会变成新的食物,哺育饥饿的生灵,或许化作新的药物,解除疾病的痛苦;有些动物也许能有效遏制作物害虫的危害;微生物或就能够分解污染物,或就能够维持住一个拥挤世界的大气。再想一想,你这一辈子能在这个迷人的新世界里的寿命是如此之短,因此你只能在这短暂的时窗内把所有的一切都发现出来。

事实上你无需穷尽你的想象力,因为这个渺茫的星球就是我们的地球。我们星球上的生命丰富得让人吃惊,它造福人类的潜能同我们幻想到的任何一个星球一样可观,更不用说我们今后还要继续求索。短暂的时窗也同样是事实。我们地球的表面,不论是陆地还是水域,都由于人口膨胀带来对住房、食物和燃料的需求而面临着急骤的退化。今后一代或两代人口的迅猛增长加之资源消耗,将在我们尚有机会发现更不用说仔细研究我们正在失去什么之前,大量生命的多样性将被毁坏。

当得知地球上数百万计的物种尚未发现或描述后,许多人都感到惊奇。实际上,地球上的生命是如此之丰富,对某些生物类群的研究是如此之肤浅,我们仍不清楚地球上到底生活有多少物种。假如地球上有一千万或五千万个物种,几乎任何一个生物学家都不会感到吃惊。

然而,我们确实知道生命多样性过去对我们有多么重要,将来对我们会有多么重要。我们也知道,研究这种多样性可以告诉我们这个星球的生命史是什么,这些研究怎样来帮助我们解决进入二十一世纪后世界面临的最紧迫的难题。我们目前正面对一个前所未有的机遇,一个在我们失败后子孙后代不再会有的机遇。

系统学是生物学的一个分支科学,它的任务是掌握生命的多样性。系统学家把掌握的知识划分到分类系统,以此表示我们对现存物种或过去年代物种的了解,进而准确推测尚未了解的物种。世界系统生物学家被迫切地要求加紧对地球生命的探索,以便为全社会提供必不可少的背景知识,发现和利用新的生物资源,并制定保护地球生物多样性的有效决策。

美国植物分类学家学会、系统生物学家学会、Willi Hennig 学会及系统学标本馆联合会组成的联合体

1994年

内容提要

地球上的物种，包括人，共同构成了一块精美的纺织品。这块纺织品塑造了对维持生命至关重要的大气、气候、土壤、水及地球的其他生态学特性。构成纺织品的一百多万根丝线（物种）为系统生物学家，即探索地球生物多样性的科学家所发现和描述。反过来，物种描述又为其他各项研究的开展打下基础，例如物种间关系的研究以及告诉我们生命多样性组成及其历史的分类研究。物种的分类犹如强大的理论工具，能帮助我们了解、维持并合理利用我们继承下来的丰富的生物财富。

系统生物学研究在过去两百多年中取得了重大的成就，尽管如此，我们对生命的了解还远远不够，数千万计的物种仍有待我们去探索、去认识。这些物种中可能有维持复杂的生态平衡的物种，可能有增加和扩大农业生产的物种，也可能有会成为消灭人类健康之敌的新药和特效药的物种。令人欣慰的是，近期系统学的发展适时提出了制订生物圈计划的艰巨任务，其目的就是在于掌握物种多样性的巨大的范围。

对人类今后的生存和幸福来说，当前迎接这一挑战比以往任何时侯都要重要。当今世界逐步面临多样性减少和栖息地消失的境况，人类对珍贵生物资源的需求也在不断提高。为此，物种多样性的基础系统学研究势在必行。世界上的自然资源管理者、药物开发者、保护生物学家、生态学家及其他有关人士只有共同努力，才能付之于实现。

对地球生物多样性了解得越多，人类保护陆地和海洋自然生境的能力就越强。如果想让子孙后代享有我们所依赖的地球生命的多样性，我们就必须求得这些知识。

国际系统生物学界倡议的2000年系统学议程旨在实现世界各国所追求的一个科学目标，即发现、描述和分类地球上的物种。实现这个目标要求国际社会努力完成三个彼此关联的科学任务。

第一项任务：全球物种多样性的发现、描述和编目；

第二项任务：分析这个全球发现计划获得的信息，并将其融合于一个能体现生命史的预测性分类系统；

第三项任务：把这个全球计划获得的信息整理成为一种有效的、可查询的形式，以最大限度地满足科学和社会的需求。

它对科学和社会的作用巨大，具体表现在：

1. 新发现的物种将扩充社会有用资源的编目；

2. 新的系统学数据将会用维持和利用各国物种多样性所必需的知识把保护学家、政策制定者及生物资源管理者武装起来。

3. 物种多样性知识将有助于新产品的发现，并将指导农作物和药物的新品种和改良品种的选择。

4. 基底数据将随之产生，并用于监测全球气候及生态系统的变化，其中包括物种灭绝速度、生态系统退化以及外来的、引起病害和虫害生物体的传播等。

导　言

"浪费、破坏我们的自然资源,……,将会损害子孙后代真正的繁荣昌盛,按理说我们应把繁荣昌盛传给他们并发扬光大"

—西奥多·罗斯福

1907年12月3日提交国会的咨文

我们与其他数百万物种共同占有地球。这些物种形态多样,关系奥秘,是三十多亿年进化的结果。地球上的物种,包括我们本身,被汇织在一个精美的生态结构中。这个结构塑造了大气、气侯及地球的地理特性,并奠定了生命本身的基础。

人类依赖于难以数计的其他物种。换言之,人类生活的好坏与全球生态网络的状况有着直接的关系。成千上万的物种被人类用于食物、住房、衣物、药品、商业或其他目的。对其他生命型的利用推动了世界经济的发展,使我们每个人的生活更富裕、更美好。

人类能有效利用其他物种的本领来自对这些物种知识的了解。这种知识从认识它们是什么种类、它们在哪儿、它们会有什么特征、它们与别的物种有什么关系开始。地球上的生命形式极其丰富多样。迄今为止,已描述的物种大概不足150万种,但据专家估计与我们一道栖居地球的物种很可能有几千万。记录和了解物种多样性对人类今后的发展至关重要。了解这些基本知识是系统学这门科学的主要任务。

系统生物学家致力于地球物种的发现、描述和了解。他们根据获得的信息对物种进行分类,在此基础上再整理有关这些物种的所有生物学知识,并制订出一个框架,预测已知的或未知的生活型的特征。尽管我们掌握地球物种的知识还不够全面,但新的系统分析方法、生物遗传物质(DNA)的直接利用、尖端的信息处理技术、自然历史标本收藏的增加及其数据库的建立等手段,都为了解全球生物多样性铺平了道路。

未加控制的人类活动使我们目前正处于一个重大的生物灭绝时期,因此目前认识生物多样性比以往任何时期更具有迫切性。据美国国家研究理事会的一份报告称,到2100年很可能一半以上的现有物种都将灭绝(NRC 1980)。但就目前生态系统的退化速度来看,这项推测可能还有些保守;按照哈佛大学著名的生物学家 E. O. Wilson 博士的谨慎估计,地球上每年导致绝灭的物种有将近2万7千种(Wilson 1992)。人口的膨胀、贫穷的增长、全球的冲突及自然资源的过度利用造成了环境质量严重的下降和物种多样性不可挽回的损失。

多样性的减少伴随着生物学知识的损失降低了各国人民改善经济状况和提高生活水平的能力。物种灭绝的悲剧向国际社会发出了严峻的挑战——在人类生物遗产永久地失去之前,发现、保存和认识其多样性。生物圈的未来尚有赖于各国政府在今后几十年内的共同努力。近期在里约热内卢召开的联合国环境与发展大会(UNCED)已经认识到这一点。大会上国际社会认识到,在保持经济持续发展的同时有必要维持生物圈的完整性。各国通过全球行动计划"二十一世纪议程"也呼吁要加强对地球生物多样性的了解程度。

在发展经济的同时兼顾持续利用取决于综合的政治决策和经济决策。成功的决策必须以各种准确的地球物种科学库的信息为依据。为迎接掌握生命多样性重任的挑战,国际系统学界提出了一项探索和研究计划——2000年系统学议程:绘制生物圈图谱。有了全社会的共同决心和一致支持,全世界的系统学家们提出了一项加速研究的计划,以期在今后25年中解答下列问题:

1. 地球上的物种是什么?
2. 它们分布在什么地方?
3. 它们具有什么样的特性?
4. 它们之间的关系如何?

本项研究获得的知识将被整理成为预测性的分类系统和数据库,使其成为认识、维持和永续利用

人类继承的巨大的生物财富的有效工具。

系统学是建立在以下工作之上的科学：

分类学：

发现、描述和划分物种或物种类群（合称分类单元）的科学

系统发育分析：

发现生物种类群间的进化关系

分类：

按照进化关系将物种最终归为不同的类群

2000年系统学议程：制订生物圈计划

世界系统学界通过2000年系统学议程果断地提出了一项迎合社会需求、有明确科学目标的计划：发现、描述和划分全世界的物种。

迎接生物多样性危机的挑战和成功地完成这项议程要求国际上的广泛参与。议程确定了三项彼此关联的研究任务：

1. 全球物种多样性的发现、描述和编目；

2. 分析这个全球发现计划获得的信息，并将其融合于一个能体现生命史的预测性分类系统；

3. 把这个全球计划获得的信息整理成为一种有效的、可查询的形式，以最大限度地满足科学和社会的需求。

系统学知识及生物多样性的价值

> “每一个国家都有三种财富，即物质财富、文化财富和生物财富。前两者我们基本了解，因为它们是日常生活的组成部分。生物多样性问题的实质是没有给予生物财富以足够的重视。这是一个重大的战略性失误。随着时间的推移，这次失误越来越会让人们感到惋惜”
>
> ——E. O. Wilson，1992年，第311页

尽管人们依赖数以万计的物种获取食物、住房、药品及其他必不可少的物质，但据科学预示生物圈的未知部分还蕴藏着更大的潜力。在全球环境面临各种不利改变的当今，掌握更多的物种多样性知识尤为重要。这些知识大部分来自系统学基础研究，即发现和描述新物种，确定物种的特征及与其他物种的关系，按照这些数据进行分类并建立预测性的信息查询系统。由此可见，对于试图了解生命多样性、为子孙后代保护和管理生命多样性的基础科学和应用科学研究人员来说，系统学知识乃基础之基础。

人类健康

世界上有几亿人在蒙受生物引发的疾病带来的痛苦。为了解除这些疾病对人类的折磨，全世界每年投入的经费高达几十亿美元。纵观世界上所有的疾病，目前最让人深恶痛绝的有三类寄生虫，它们是导致疟疾的原生动物、引发日本血吸虫病的血液蠕虫及造成爱滋病的HIV病毒。此外，世界上25%的人在肠内寄生有影响儿童身体和智力发育的蛔虫。自1979年以来，细菌中有270个新属和大约1100个新种被相继描述。现已认识的病原或兼性病原属、种的数量在不断增加。

没有系统学这门科学，就不可能取得对付这些疾病的进展。系统学家能够认识、区分并说明影响人类健康的非病原体及病原体的特征。这些生物有数十万个种和上百万个品系，其中包括细菌、病毒、真菌、酵母菌、原生动物、线虫、蛔虫、扁虫、绦虫、昆虫、蜱、螨、蜘蛛、蝎、蜗牛等。仅仅了解可能引发疾病的生物这还不够，我们还要把一个类群中的所有已知物种区别开来，以便加强对新发现物种的了解，确定某个已知的品系或物种本为非病原体是否已转变成病原体（见专栏1）。

掌握致病生物的进化关系对提高人类健康水平同样起着关键的作用。通过研究疾病载体及相关的非疾病载体的相似性，便能预测致病生物的变化趋势和发现新的病原型。由于爱滋病导致的免疫紊乱患者人数不断增加，而人类又能够移植器官、医治重度烧伤和延长老人及癌症患者的生命，因而，即使是最无毒的细菌和病毒也有可能危及生命的安全。尽管人们有抗生素、疫苗、良好的卫生条件和安全的食品，但是许多传统的疾病正在复活，一些前所未有的疾病正在出现。大量曾一度被认为是非病原体的物种，目前已经从人类病例中被分离出来。事实证明，掌握全部生物类群的进化关系及地理分布知识，对于了解病原体由动物转至人类的过程和发现某些病原种比其他种毒性更强的原因起着关键的作用。

"如果认为我们必须在服侍人类和服侍环境之间作出选择的话，那我们就犯了一个危险的错误。两个目标的统一必须当作一个首要问题来对待，二者不能也绝对不能分离开来。"

—Orville Freeman，
美国前农业部长，1989年

物种经济学

物种的利用给全球经济带来了价值数万亿美元的效益。随着越来越多的国家编目、研究并提出对本国境内多样性的权利时，关于物种的利用和管理的国际协议和条约逐渐在受到重视。发现和描述的物种越多，对其分布及与别的物种间的关系了解得越详尽，那么这些物种对一个国家的经济能作出的贡献就越大，这些物种能给后代保存下来的就越多。2000年系统学议程所确定的研究任务已经为获得这些重要的知识制订了计划。

历史表明，新种的发现以及随后的特性研究，往往会带来重大的经济效益。系统学分析，包括物种间已研究认识的特性的比较，从而能够推测新种的特性；反过来，这些推测又能更有效、更可靠地评定这些新种的潜在经济价值。

药　　物

据世界卫生组织统计，人类为药用目的而利用的植物有两万多种。事实上，发展中国家有80%的人仍然依靠传统药物作为主要的疾病治疗（手段）。这些药物大部分来自野外采集的植物，给这些物种的很多野生种群带来了很大的压力。

专栏1

疟疾与系统学：既救命又省钱

没有遵循寄生虫及其传播媒介的准确的系统学知识而制定的疾病控制措施，只会造成时间和金钱的浪费。疟疾就是一例。

疟疾是由 *Plasmodium* 属的寄生原生动物引起的。在该属四个会侵染人体的物种中，*Plasmodium falciparum* 造成死亡或病态的危害最大，它通过蚊子叮咬而传染给脊椎动物。能携带寄生虫的蚊子的种类很多，但它们的分布和传播能力各不相同。据世界卫生组织估计，全世界疟疾患者每年有2—3亿例，大约一百万例患者因病死亡，其中多数为儿童。

六十年代，许多科学家都在从事 *Anopheles gambiae* 这种分布于非洲的主要传播媒介的杀虫剂抗性研究。几十个品系被送到伦敦进行杀虫剂和杂交试验，以确定其抗性遗传模式。大量的杂交种不能再生育。在研究过程中，系统学家发现一个被认为是 *A. gambiae* 的种实际上是六个不同的种，其生物学特性和疟疾传播能力显然各不相同。类似情况在蚊子中还重复出现过。美国的 *Anopheles quadrimaculatus*、印度的 *A. culicifacies* 和泰国的 *A. dirus* 均被证明是种的复合体（即不仅仅是一个物种），它们传播疟疾的能力各不相同。

"美国药店开出的所有药方中，四分之一是从植物中提取的物质，13%来自微生物，另外3%强来自动物。也就是说，40%以上的药来源于生物。然而，这些药物只不过占众多可利用中的极小部分。"

—E. O. Wilson，1992年，第283—285页

实践证明，在筛选有药效的成分时，如果选用传统药物所利用的物种，其成功机遇比选用同一地区中植物随机抽样要大得多。人类利用动物和植物的能力，是各地人民与自然界密切相处，在长期的实践、失败过程中经过不断总结形成的。专长于民族生物学，即研究人们利用土生土长动植物的系统学家对这些重要经济物种的描述和研究起着主要作用。开展人类利用的动物、植物和微生物的民族生物学综合调查迫在眉睫。记载传统人类社会与为其所用的动植物间的关系，包括种质保护，是系统民族生物学的一个重要研究领域。

分类系统的创造是系统民族生物学的一个最重要的贡献。系统民族生物学的民族生物区系遗传

多样性的大部分科研工作，如保存、保护、育种、有用基因的探索、生物技术等，都随着近缘关系的决定而定的。研究成果多以分类、分类学专著或系统发育分析的形式在科学界交流。这些研究可提供基础的生物学信息，如物种分布、结构变异等，它们对驯化种和野生近缘种的研究有着关键的作用。许多发现新药的例子也可以说明系统民族生物学的重要性(见专栏2)。

世界上用有花植物制造的药物产值达数十亿美元。然而这项财富大多数仅来自一小部分物种。与25万种有花植物相比，微生物具有更大的多样性，它们很可能是药物和其他生物技术产品的重要来源(Bull 等，1992)。微生物药中最著名的例子是由 *Penicillium notatum* 生产的盘尼西林，它给治疗传染病的药物带来了一场革命。

寻找微生物来源新药的工作才刚开展。微生物的系统发育和分类能促进对它们相互间关系的认识，这为筛选自然界的大量物种提供了可预测的线路图。因此，系统学知识对微生物新药的发现具有不可估量的作用。遗憾的是，人们对微生物尤其是病毒、细菌和真菌的系统学了解甚少，上百万的物种仍有待发现和描述，多数类群间的关系仍有待分析。此外，研究许多微生物类群的系统学家也不断减少。这种状况如不改变，人类便会失去发展经济和技术的许多良机(Hawksworth 和 Ritchie 1993)。

专栏2

系统民族生物学与新药的发现：马达加斯加的几个例子

马达加斯加这个陆地岛屿上有1万3千多种植物。令人惊奇的是，这个岛屿上80%的植物都是特有种。通过研究这些特有种及其在当地医疗体系中的作用而开发的新药不胜枚举。

长春花(*Catharanthus roseus*)是其中最著名的一个例子。不久以前，它还仅仅是人们花园中的一种观赏植物。当地不少人把它当作一种民间草药来治疗糖尿病，于是人们对它进一步研究，以期从中找到口服胰岛素的代用品。虽然玫瑰红蔓长春花在治疗糖尿病方面的效用尚未得到证实，但它的提取物却发现具有大幅度减少实验动物白血球计数和抑制骨髓活动的功效。根据上述观察结果，人们终于分离出两种化学成份，即能够有效治疗白血病的长春碱和长春新碱。自从这种药物首次进入市场以来，儿童白血病的治愈率由过去的10%提高到现在的95%。

芙木属(*Rauwolfia*)中有两个种，即蛇根木(*R. serpentina*)和 *R. vomitoria*，它们的根是提取几种生物碱或制作根粉的原材料。这些产品的药物制剂可用来治疗高血压，也可用作治疗精神紊乱的镇静剂。

"……问题是灭绝的速度上升得如此之快，以至于如果人们找到一种具有特殊生物活性的植物，很可能当你去时已发现其生境全然无存。这是一场时间战争。"

Michael Balick

纽约植物园经济植物研究所所长

农　　业

世界农业的发展有赖于农业研究带来的技术进步。从世界范围看，发达的农业体系正朝着减少杀虫剂、化肥和除草剂的施用量，加强生物防治、害虫综合管理和持续性农业的方向发展。以上技术强烈依赖于有关害虫类群、害虫的植物寄主及害虫天敌的系统学知识。系统学信息是农业管理的语言和预测基础，由于系统学信息的不足而造成搁浅或失败的项目多不胜数(见专栏3)。

随着对非杀虫剂防治策略的作用的不断重视，需要了解在农业生态系统中起重要作用的各种各样的生物越来越成为关键问题。据估计，有用的生物防治因子可能有数千种，但它们均未被科学所揭示。因此，要发挥它们的经济作用，就必须首先发现和描述这些生物，并把它们纳入分类系统和信息系统。如果有一些生物，它们在农业生态系统中既能使产量提高并又有一个健康的环境，但人们却对它们尚未认识，或与别的生物混淆不清，或与其他物种的关系不明，农业发展势必会受到极大的阻碍。

专栏3

系统学研究节省数十亿美元：几个生物防治事例

在十九世纪末期，吹棉蚧(*cottony-cushion scale*)严重影响了加利福尼亚的柑桔产业。根据一位系统学家提供的信息，在澳大利亚进行国外考察，结果发现并引进了一种以吹棉蚧为食的瓢虫，从而控制了这种蚧虫的危害，拯救了加利福尼亚的柑桔业。

许多年来，生物防治专家一直没有找到对加利福尼亚红介壳虫有效的天敌。一位介壳虫专家应邀研究这种昆虫，结果发现它实际上是三个不同的相似种，即加利福尼亚红介壳虫、黄蚧和紫杉蚧。其后，研究寄生蜂的系统学家发现，正是因为鉴定失误才难以找到红介壳虫的天敌。不久，各种成功的生物防治因子相继被引入。

了解有害种的原产地往往会给寻找有效生物防治因子提供很大的帮助。甜菜叶蝉(*Circulifer tenellus*)早先被认为是 *Eutettix* 属的一个种，原产南美洲。遗憾的是，人们在南美洲不仅没有发现防治因子，还浪费了大量的时间和精力。最后系统学家发现该种其实属于东半球的 *Circulifer* 属，于是便在地中海地区找到了几种有效的天敌，并引入加利福尼亚。

1974年，扎伊尔发现了一种引进的粉蚧科介壳虫。这种害虫每年给西非的木薯种植业造成近14亿美元的损失。这种介壳虫被描述为物种 *Phenacoccus manihoti*，于是人们开始在南美北部寻找它的生物防治因子。在没有找到任何有效的寄生虫后，一位粉蚧科系统学家又重新研究，结果找到了一个与上述种非常相近的种 *P. herreni*，主要见于南美北部，而 *P. manihoti* 实际上分布在较南部。有了这一发现，有效的寄生虫便很快被找到并引入非洲灾区。

外来害虫的侵入对世界农业生产最具破坏力。不幸的是，国际贸易量的增长和交通网络的飞速发展增加了引入新害虫的可能性。大多数国家都要查验入境的商品，搜查任何外来的传染物，并制定相应的入境管理规定。这些措施基本上依赖于系统学家提供的信息，因为任何一个国家都不可能有完整的所有物种的编目，也不可能完全了解某种被查获的生物是否分布于本国。再者，多数重要的农业类群大都确乏详尽的鉴定材料。缺乏某一害虫类群的系统学信息有时还会产生一些特殊问题，例如，烟草芽蠕虫(budworm)往往见于来自中美和南美的查获商品。芽蠕虫一直被认为只有三个种，但近期专项研究表明，其复合体由十二个不同的型组成。

各个国家也必须制定若干规定，以限制从某些国家进口某些特殊的农业产品。这些规定都必须以系统学知识为依据，同时也要依靠本国对原产国商品中出现的所有植食性生物、它们是否有害虫的潜势以及世界分布的了解程度(见专栏4)。

由于缺乏足够的系统学知识，尤其是关于有害物种的系统学知识，制定这些重要规定的科学依据往往不足。

专栏4

系统学信息与农产品贸易

从某种情况上讲，掌握有潜在危害的生物的系统学知识可为增加商品的出口量和避免严重国际贸易事件发生铺平道路。举例说明，加拿大从美国进口的小麦中带有一种黑粉菌，这种菌最初被鉴定为 *Neovossia indica*，一种极其危险的外来病原体。为此，加拿大作出禁止从美国进口小麦的禁令。经一位系统学家研究这种病菌，最终认定是一种常见于大米的物种 *Tilletia barclayana*，并且早在小麦运到同一仓库之前就已存在。正是由于这项极小的系统学发现，一场可能造成重大经济损失的国际事件才得以平息。

农业和遗传资源

在过去，农业项目提倡使用杀虫剂和化肥，常常给环境带来不利影响。而在当今，随着害虫管理和生物防治等低环境影响策略的出台，作物种质需要重新发掘。这项工作要求详细地了解世界珍贵的农

业生物区系，譬如说潜在的作物体系、害虫、控制害虫种群水平的天敌等。遗传工程能成功地把一种生物的基因移植到另一种生物，因此，保持物种多样性变得尤其重要，因为任何生物的遗传材料都可能有益于农业体系。系统学知识则能使科学家更好地了解作物的抗病情况，从而在特定情况下选用最佳的遗传原种(见专栏5)。

专栏5

系统学发现与作物改良：两个事例

玉米

1977年，一位叫 Rafael Guzman 的墨西哥植物系统学家重新发现了(在两块玉米大田间的水渠中)消失已久的 *Zea perennis*，一种珍贵的多年生玉米草，即野玉米。*Zea perennis* 是人类三种最主要农作物(其他两种为水稻和小麦)之一的玉米(Zea mays)的一个近缘种。不久以后，Guzman 又在附近山区的云雾林中发现了这种多年生玉米草的另外一个种，并把种籽送给玉米系统学家 Hugh Iltis 和 John Doebley 研究。这些种籽所产生的植物结果是一个鲜为人知的种 *Z. diploperennis*。与 *Z. perennis* 不同的是，这种玉米的染色体数目与家种玉米恰好相同，很容易与家种玉米杂交。更令人感到惊奇的是，*Z. diploperennis* 可以抵抗家种玉米(*Z. mays*)常见的七种病毒病，同时某些对病毒的抗性目前可以移植到家种玉米。迄今为止，已经有四个抗病毒玉米品种进行了商业性种植。世界玉米的年产值近600亿美元，这项发现的潜在经济价值可见一斑。

上述事例的另一方面也值得引起人们的重视。这个新种发现于生物多样性丰富的墨西哥西南部地区，但是森林砍伐、牧业和农业正在迅速破坏着这个地区的森林。如果不是 Sierra de Manantlan 整个山区被确定为生物圈保护区，这种珍贵植物所剩的几个种群现在可能已经绝灭。还有，新种的发现者如果不是一位杰出的草本植物系统学家，或者他没有把种籽送给分类困难的 *Zea* 属方面的专家，这一珍稀物种或许就会为人们所忽略。以上事实进一步说明了培养一大批优秀的系统学家，增强探索、记录和保存世界上现存生物区系多样性力量的重要作用。

蕃茄

1962年，Hugh Iltis 和 Don Ugent 在安第斯山探险时采集到一千多号植物标本，包括一个野生番茄新种的种籽。新种与家种番茄杂交后，可增加杂种果实的可溶性固体含量。这项成果每年可为番茄种植业创造800万美元的效益。

林　　业

美国的林地面积有230万英亩。虽然木材产品的产值高达1360亿美元，与其相比，林副产品及娱乐、水资源等非商品利用所创造的价值则要更大些。随着人们对森林价值认识的转变，森林管理出现了重大的变革。新型的管理哲学促使人们去保护老龄林，并寄更大希望于非木材物种。与传统的森林管理方式相比，森林的长期管理面临着一连串的问题。尽管大多数森林管理者能够测定森林物理和化学性状，监测动植物区系变化，但他们仍然缺乏衡量森林系统生物多样性的专门知识。有限的无脊椎动物及微生物系统学知识使森林濒临险境，在控制引进的有害节肢动物和病原体方面尤其如此。例如，1992年不列颠哥伦比亚发现了一种奇异的昆虫。由于没有这方面的专家，等鉴定出这种可怕的外来甲虫 *Buptestis hemmoradialis* 时已经过了一年的时间。在此期间，这种昆虫已经过了两个生长季节。这么一来，本来可以在小范围内采取的灭杀措施，只好扩展到大范围内进行。像这样的时间耽搁不仅降低了灭虫效果，而且还增加了灭虫的经费开支。

渔　　业

渔业产品是世界上蛋白质的主要来源(Norse 1993)。因此，区分普通鱼类和具商业价值的海味对自然资源管理和为水产养殖物种的选择有着重要的作用(见专栏6)。此外系统学信息在国内法及国际法规、条约和公约的执行等政策问题上也同等重要。

非本地种的引进极有可能给水生生境带来严重的危害。人们往往认为一些外来种有益而故意将其引入，不料后果最为严重。不少水生物种通过栖居船体或水舱而无意被引入。这些引进的物种大部分会在经济上给土生种带来严重的损失。再则，引进种还会携带许多寄生虫和病原体。系统学的研究对准确鉴定引进种及其寄生虫和病原体有着不可磨灭的贡献，是制定有效管理策略的理论基础。

专栏6

系统学研究提高渔业产量：两个事例

正确鉴定目标物种把系统学与渔业也联系到了一起。太平洋东北部的狭鳕（*Theragra chalcogramma*）捕捞业是世界上最大、创造经济价值最高的捕捞对象之一。至二十世纪八十年代，狭鳕已经成为世界渔业中捕捞量最大的鱼种。然而，狭鳕只不过是东北太平洋地区几种近缘鱼类中的一种，不久前其早期生活史还鲜为人知，部分原因在于其幼体的野外样品难与其他鱼类难以辨别。在过去的十年中，系统学的研究结果突破了这一难关，能够准确地分辨出不同的幼体，使得通过增加幼体数量提高可捕捞鱼类产量之计划得以实施。系统学家还通过随船服务和举办管理人员分类培训班，为这项捕捞业提供后续服务。

要维持和扩大鱼类可捕捞种群的规模，准确无误的系统学信息必不可少。例如，有关西班牙鲭的数据都以巴西种群为根据，渔业人员最初打算利用这一信息来管理墨西哥湾和美国东海岸的种群。然而系统学研究证明，巴西种群为 *Scomberomorus brasiliensis*，与北美种群 *Scomberomorus maculatus* 截然不同。由于两个种的生物学特性各异，像南方种那样进行管理很可能会导致失败。

了解和保护地球的生命支持系统

地球表面的环境与有生命的生物间的关系十分密切，并随时间的推移而改变。地球上数百万的物种不仅相互关联，与周围的环境也息息相关，最终形成了一个维持生命存在的错综复杂的生态网络。这种关系的产物免费为我们提供了清洁的空气、水、肥沃的土壤及调节地球的地球化学循环（基本上通过微生物作用完成）。绿色植物能吸收太阳的能量，并把能量转化给别的生物。世界上的植被特别是热带雨林再把水循环到大气，从而控制着气候的变化。

随着人口的暴涨及全球变化的加快，地球的生命支持系统日益受到威胁。人类从世界自然资源获取食物、住房、衣物和燃料的同时，也给环境带来了严重的影响，像森林大面积砍伐、空气污染、水污染、全球变暖等。

系统学的知识在监测这种全球变化方面起到最基本的作用。收藏的标本是生物群落和生态系统变更的见证，记载着长时间内环境对所受压力的反应。同样是这些标本，由于包含不同的物种存在和鉴定的基本科学证据，因此也是物种灭绝的最可靠记载。对下个世纪物种灭绝的推测，主要依靠森林砍伐和生境破坏方面的综合信息。没有物种存在和分布的有完整记录的科学知识，就没有生态变化和物种灭绝的准确评估。只有系统学才能可靠地衡量生物多样性的危机程度。

另一方面，正确鉴定物种对监测全球变化也十分重要。所有的生物群落都包含一些对环境变化特别敏感的物种。比如，某些蛙类对空气质量的变化异常敏感；在水生群落中，某些鱼类对水纯度变化十分敏感。为此，科学家已越来越多地利用这些指示种来考察全球变化对自然群落的影响，从而实现监测全球变化的目的。只有准确鉴定和描述这些物种，掌握其分布及近缘种的知识，才能够开展这些监测活动。

世界上许多生境和生态系统都生活着成千上万的有着极其复杂作用关系的物种。生态学家和资源管理人员进行这些相互关系的动态研究时，由于对即使是最普通物种的鉴定和分布的认识都还存在着缺陷，这些生境和生态系统的基本描述当然不可能全面。这就要求开展深入的系统学研究，描述和鉴定地球众多生态群落中生存的物种。这些信息对提供评定环境压力所依据的基低数据至关重要。

系统学对自然资源的管理和保护同样有重要作用。保护区生物多样性的保护管理人员需要对物种作鉴定并了解其地理分布，为有效管理策略的制定和实施提供依据。系统学则为物种的鉴定、多样性的评价以及需要特别保护的物种的确定提供了理论基础。此外，系统学信息与保护区和开发区的选

址和规划、它们与地方和国家法规的关系的评估，都有着紧密的联系。有效管理动植物国际贸易也同样需要精确的系统学资料。这些资料还直接有利于一些像濒危野生动植物种国际贸易公约(CITES)的国际性条约和公约的执行和实施。

提高日常生活的质量

生物的多样性有助于我们去理解人类意识和智力的某些方面。对自然环境的宁静和舒适的追求是人类的共性，这就是许多宗教道德文化为什么具有保护和尊敬自然环境信仰的原因。人类能够使其他物种灭绝，而在伦理道德上，人们普遍认为应肩负起阻止这种悲剧发生的责任。这正像 Ehrlich 和 Ehrlich(1992，第220页)指出的那样，“我们认为，如果大多数人没有保护生物多样性的观念，那么这个问题就不可能得到解决。”

纵观人类历史，人对其他物种显示出极度的好奇心，并为其美学价值所倾倒。人类与自然的这种联系在所有民族中普遍存在，并体现于对其他物种的态度，尤其是园艺业、宠物饲养、野生动物和鸟的观赏等活动之中。对其它物种的美学和情感依托的同时，还为许多人创造了巨大的经济效益，尤其在自然旅游和标本贸易这两个方面。

旅游业每年创造的价值达2500亿美元。如果把观赏其他物种连带的旅行活动计入在内，上述收益中有将近20%来源于生态或自然旅游。不少国家国民生产总值的大部分或部分来自生态旅游，其中单是某些物种就创造了巨额收入。例如在东非，肯尼亚 Amboseli 公园的一头狮子十五年内可创汇50万美元，所有的非洲象每年创汇60万美元。如果计入此类自然区域各项配套产业的盈利，这项收入在全世界可达几百亿美元(见专栏7)。

专栏7

系统学与生态旅游

系统学研究对生态旅游业的贡献重大。系统学方面的出版物，例如修订版、图解、专著、编目、标本等，为野外向导、旅行向导、电影、录像、录音及其他宣传手段提供了科学的背景材料。

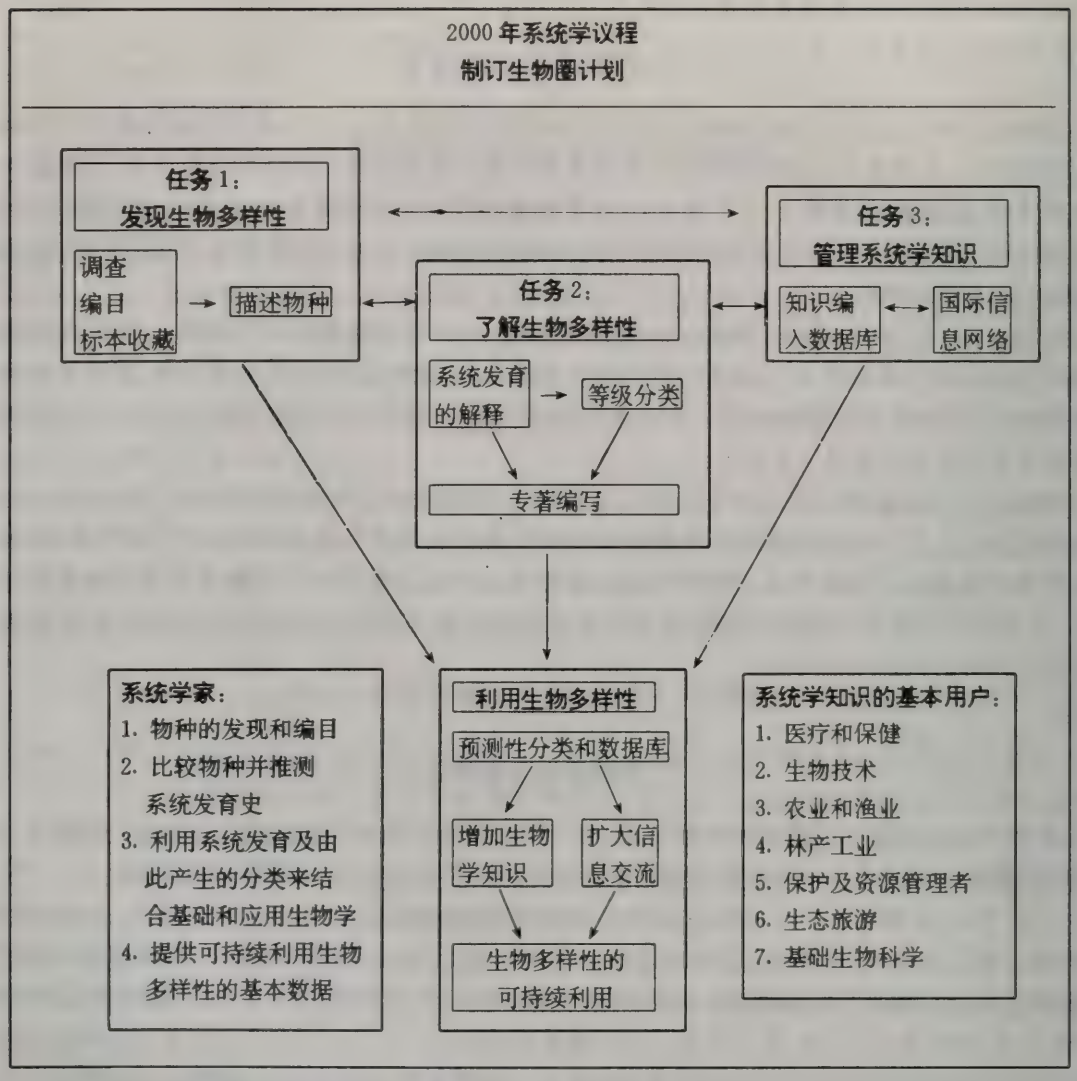
上千年来，人类不论生存在哪里，都始终和其他物种的驯化、种植打着交道。当今驯养和栽培的动植物已有几万种，其繁殖和贸易的商业价值有几十亿美元(Groombridge 1992)。举例说明，仅英国就栽培有3000多种用作装饰的植物；全世界有贸易记录的兰花种类超过5000种。从国际商业范畴来讲，像鸟类、爬行动物、蛙类、热带鱼类、蝴蝶、蜘蛛等动物的出口每年可创造几亿美元的效益。系统学对监测和管理世界自然资源的重大作用表现在它提供了准确的识别材料、图解以及物种分布信息。世界上大量的动植物贸易对其自然种群产生了极大的压力，尤其是从野外直接猎取。由于许多物种形态的相似，因此，有效地监测物种和执行法律必须依靠物种的准确鉴别。

加强科学研究

系统学构画了一个总体框架，使生物学的研究成果在这个框架内得以整理和交流。比较两个或两个以上物种的生物学研究，或针对某一物种但其研究成果最终会被对其它种有兴趣的生物学家考虑的生物学研究，都或多或少地借鉴系统学的研究结果。系统学研究的结果对选择研究系统和评价一些有意义的生物学现象的普遍性也有重要的作用。

通过分析物种关系而产生的系统发育树包含着对血统(共同祖先)、长期的特征变化及地理分布(历史生物地理学)变化的假设。理论上讲，在演化世代期所发生的所有生物学变化都能在系统发育树中体现出来。近年来，由系统学家提出来的以系统发育假设为依据的严谨的比较研究法得到了发展，用来研究和解释一系列生物学现象，例如寄生物及寄主的生物地理学和协同进化，生态学和行为学的历史变迁等(Brooks 和 McClennan 1991)。与此同时，系统发育假设也成为研究适应、物种形成、灭绝等基本进化过程的理论基础。

2000 年系统学议程的任务



第一项任务：全球物种多样性的发现、描述和编目

生物圈物种的发现、描述和编目是人类用智慧管理生物圈的一个重要贡献。由于许多物种过于微小，其研究十分困难，造成世界各地生物多样性的了解甚少。譬如说，世界上任何一处普通庭院都生存有许多物种，将这些物种逐一列出几乎不大可能，更何况比它复杂得多的生态系统。

迄今为止，系统学家大约已经描述了140万种生物，其中大部分是昆虫。据推测，尚未发现和描述的物种仍有1千万到一亿种(见专栏8)。

有待了解的物种的数量多得如此令人吃惊，这似乎又与我们在日常生活中所接触到的少数树木、哺乳动物、鸟类、蝴蝶及其他常见物种大有出入。那么就去想象一下大片的雨林，那里仅在一棵树上就能发现几百种昆虫，在一小块森林就生存有几百个不同的树种，仅在几立方英尺的土壤和地表腐叶中就会生活着上千种微小的螨、线虫、真菌和微生物。

专栏8

物种数量知多少?几个事例

	已描述物种的数量	有待发现物种的估量
病毒	5千	约50万
细菌	4千	40—300万
真菌	7万	100—150万
原生动物	4万	10—20万
藻类	4万	20万—1千万
植物	25万	30—50万
脊椎动物	4.5万	5万
线虫	1.5万	50—100万
软体动物	7万	20万
甲壳动物	4万	15万
螨、蜘蛛	7.5万	75—100万
昆虫	95万	800万—1亿

引自 Groombridge(1992)

人类对全球环境造成的压力,迫切要求尽快掌握世界生物多样性的知识。这项工作要求我们加大探索地球的力度,汇集这些物种的样本,仔细分析发现的标本,以准确判断它们与已知物种的区别。要实现此项目的,就必须支持对地球生境开展综合性的考察和编目以及对收藏标本的研究工作。今后的几十年中,地球五分之一之多的物种行将灭绝,但目前对全球生物资源的编目远远不够,难以提供物种多样性的信息。物种多样性充满活力,可推动世界走向一个可持续发展的未来。

本项任务中,应优先考虑以下几个方面的工作:

1. 调查海洋、陆地和淡水生态系统,获得全球物种多样性的综合性知识;
2. 确定这些物种的地理分布和时间分布;
3. 发现、描述和编目受威胁及濒危生态系统中生存的物种;
4. 把了解最少的生物类群作为目标;
5. 对能够维持全世界各种生态系统的功能及完整性、促进人类健康、改善人类食物来源的关键类群进行编目。

全球物种编目的作用

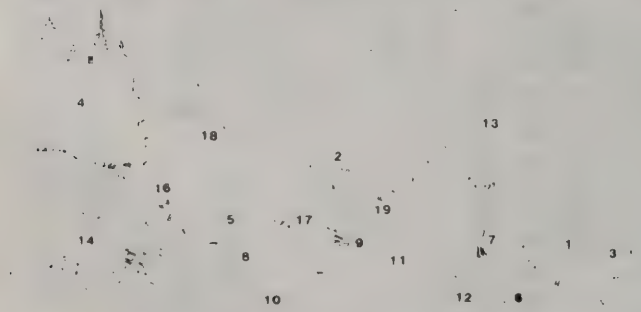
1. 能推动物种的发现和分类,综合有关物种的信息并编入数据库,便于快速查询。
2. 能获得世界上很多生态系统内物种多样性、分布、特性方面的信息。
3. 能进行本底评估,对全球变化进行长期监测和分析。
4. 能发现新的生物资源。

基于科学、经济和伦理上的原因,我们必须阐述和了解物种多样性的重要性,以免为之过晚。本项任务的效益远远大于迎接生物多样性危机之挑战所需要的投资数额。我们没有比这更大的科学举措,也没有比这更好的机会。

第二项任务:分析这个全球发现计划获得的信息,并将其融合于一个能反映生命史的预测性分类系统

2000年系统学议程的第一项研究任务包含着一个生物多样性最基本的问题:哪些物种与我们共同栖息地球?除此之外,系统学家还有另外一个研究目标,这就是记录物种的特性,综合生物学其他领

物 种 图



各种生物的大小代表其主要分类单元已描述物种的数量。
单位尺寸:□ = 约 1 千个已描述的种

分类单元

1 原核生物界(细菌、蓝绿藻)
2 真菌
3 藻类
4 植物界(多细胞植物)
5 原生动物门
6 多孔动物门(海绵)
7 腔肠动物门
8 扁形动物门(扁虫)
9 线形动物门(线虫)
10 环节动物门(蚯蚓等)
11 软体动物门(软体动物)
12 棘皮动物门(海星等)
13 昆虫纲
14 非昆虫节肢动物门(螨、蜘蛛、甲壳动物等)
15 鱼纲(鱼类)
16 两栖纲(两栖动物)
17 爬行纲(爬行动物)
18 鸟纲(鸟类)
19 哺乳类(哺乳动物)

插图:Frances L. Fawcett, 作者:Q. D. Wheeler, 1990, 美国昆虫学会年刊 83:1031 - 1047。

物种图是一幅想象的景观。图上生物的大小与其代表类群的物种数量成正比。这些数量还不够准确,许多物种尚未被发现或描述,对大部分物种的系统发育关系也了解甚少。例如昆虫(甲虫)以迄今描述的 95 万种为准,但据昆虫学家估计可能有 1 千万种以上。图中涉及的其他类群如爬行动物和许多无脊椎动物的分类单元,可能未按其自然类群组合,也就是说它们可能不具有一个独特的共同祖先。随着物种的描述、其间关系的分析和分类对系统发育的反映,本图将全面、形象地体现生物多样性。

域的数据，并为解释生物学信息提出一个理论框架。仅仅具备一份物种的名录并不能得到整理所掌握知识的一个预测性框架。永续利用地球的生物多样性，预测的准确性十分重要。事实上，系统学家已经掌握了一套整理生物多样性知识的科学依据，其中包括通过了解物种间血缘关系或系统发育关系而产生的分类系统（见专栏9）。

现生物种是进化多样化漫长历史的最终产物。进化史上共同祖先和相关性之独特模式是建立物种系统发育的基础。按照这种系统发育遗产进行的分类称为自然分类。

物种的知识是按照生物系统发育分类进行整理的。获得这些知识能使科学家们推断有特殊科学价值或社会价值的物种及其特性，并有利于创造保存这些知识的有效的信息系统（见专栏10）。这些知识还将推动新的多学科研究。按照自然分类的理论框架储存和查询信息，能够更有效更经济地利用物种及其生境的知识（见专栏11）。

专栏9

系统发育知识使基础生物学与应用生物学相结合

分析系统发育关系是一项综合性的工作，可把它比作是汇集各研究领域数据的一个总成。这些研究领域包括大解剖和显微解剖学、发生生物学、遗传学、分子生物学、比较生理学、地质学、古生物学、生态学、行为学以及生物地理学。一项研究中最后采用的数据取决于所研究生物的种类、与这些生物有关的知识的多少以及不同类型的数据对确定种间关系所起作用的大小。

系统学家把各种生物特征的相似性归结到等级模式之中，以反映生物的系统发育史。这样，猫类物种具有的相似性在熊类物种中则没有；像猫类和熊类的食肉动物所具备的相似性，其他哺乳动物中也不存在。更广泛一些，所有哺乳动物具有的特征，如毛皮和乳腺等，其他脊椎动物都没有，依此类推，直到包括所有生命类群。由此，掌握这种历史等级使系统学家能够准确地预测没有仔细研究过的物种的特征。这种推测有着重大的经济意义。从鲜为人知的物种中寻找新的活性药物化合物就是其中一例（见专栏11）。

专栏10

自然分类如何进行预测

正像物种特征比较研究可以揭示种间进化关系那样，进化关系也可用所谓的自然分类清楚地表示出来。物种间的关系划有等级，其分类也是如此。这种关系按照十八世纪瑞典分类学家卡尔·林奈的姓名命名，称为林奈等级，包括种、属、科、目、纲、门和界。照此分类，所有哺乳动物都属于哺乳纲，其中包括食肉目。食肉目又包括熊类和猫类，分别属于熊科和猫科。猫类被划分在猫科猫属 *Panthera*（注：仅一部分猫类划分在这个属），这一属内还包括虎，它的学名（拉丁名）为 *Panthera tigris*。这些独特的名称好比一种科学的语言，不仅便于交流，而且也便于储存任何生物类群的信息。

由此可想而知，按照这种等级列出的物种名单比按照字母顺序列出的名单更能预测物种的特征。因此，一种了解甚少的猫科动物可以预测它具备其他猫科动物的许多特征、食肉动物的一些特征和所有哺乳动物的部分特征。按照字母顺序列出的名单，或没有准确反映种间关系的等级分类名单，则不具备这种推测功能。

上述简单事例在人类应用物种知识方面具有深远的意义。假如为增强对物种生物学特性的预测能力，把这种知识聚集在数据库内，从此便能更有效、更经济地利用这些特性的话，我们就会在自然分类的理论框架内储存和查询这些知识。这进一步突出了发现所有物种系统发育关系的重要性。

专栏11

分类的预测价值：紫杉醇与癌症

天然产品紫杉醇产自短叶紫杉(*Taxus brevifolia*)的皮，经证实是一种治疗卵巢癌和乳腺癌的强效药剂。遗憾的是，从三棵树的皮中提取的紫杉醇才能治疗一位癌症患者，而且事后树便死亡。漫无目的地去寻找含有同样产品的植物可能会花费很多年的时间。由于了解短叶紫杉的进化关系，人们便着手审查它的近缘种。研究发现，少量浆果紫杉(*Taxus baccata*)的叶也能够提取紫杉醇，而且既经济又不会给紫杉树带来危害。

更有趣的是，发现一种生活在短叶紫杉皮中的真菌新种 *Taxomyces andreanae*：它也能产生紫杉醇。这样便开辟了生产廉价、有效抗癌药的可能性。

过去的二十年极大地推动了系统学理论和实践的快速发展，使揭示地球生命的多样性和系统发育成为一个可以实现的目标。电子显微镜术、基因序列分析等新的数据收集方法进一步拓宽了信息来源，反过来，信息又协助人类把物种划入等级分类，使信息本身成为所有比较生物学领域内整理物种多样性知识不可缺少的框架。新型计算机技术能够进行大型数据库的处理，否则至少上一代系统学家会被埋没在大量的数据之中。

本项任务中，应优先考虑以下几个方面的工作：

1. 确定主要生物类群的系统发育关系，为基础生物学和应用生物学提出一个总体理论框架；

2. 发现对应用生物学至关重要的物种类群的系统发育关系，重点放在对人类健康、粮食生产、全世界各种生态系统保护等有重要作用的物种上；

3. 发现对基础生物科学至关重要的物种类群的系统发育关系，比如与实验科学有广泛联系的学科，或者对维持生态系统功能及完整性必不可少的学科；

4. 继续探索对分析系统学数据更有效的技术和方法。

生命系统发育分类的作用有：

1. 建立一套测定灭绝速率和全球变化模式的框架；
2. 为寻找基因、生物产品、生物防治因子和潜在作物物种提供指导；
3. 为生物学知识的管理提供一套预测性框架，为社会和科学的沟通奠定基础；
4. 协助保护者、资源管理者和政策制定者确定优先项目；
5. 为连结所有生物学科的比较研究、连结所有生物类群的多学科研究创造基础；
6. 为了解形成当今生命多样性的物种形成、灭绝、适应等过程提供科学基础。

第三项任务：把这个全球计划获得的信息整理成为一种有效的、可查询的形式，以最大限度地满足科学和社会的需求

掌握世界数百万物种的大量知识，就需要新型的配套系统的支持，这样才能有效地利用和查询这些知识。这种信息系统包括系统学、地质学和生态学方面的数据，其来源途径主要是现有自然历史标本收藏、图书馆、档案馆以及不断发展着的调查和编目。随着新种的发现、系统发育关系的阐述以及其他信息的积累，有关物种的知识也需要不断更新。电子知识在全世界的普及使所有国家都受益匪浅。数据库例如分子遗传学数据库，必须通过标准的物种名称和系统发育分类与其他生物技术数据库相连接。

系统学的发展依靠物种信息的积累和交流。在过去，信息是通过植物志、动物志、专著、民族生物学研究等文字印刷方式进行积累和传播的。除此之外，还有几亿个标本和与标本有关的数据散存在世界各地的系统学标本存储机构。这些信息必须按照科学名称进行整理，同时科学名称要遵循以种间关系为依据的等级分类。

高速的信息处理技术、硬件以及建立有亲缘关系的数据库等现代化信息手段的出现，使人们能够为任何意图、以任意组合方式提取物种信息。在综合其他数据库内的信息后，将为提出地球生命组合

新的见解提供依据。另一方面，以上的数据库还能避免研究和管理的重复，节省数百万美元的经费。然而，由于各种来源的欠缺，彻底改革系统学在实际应用上的作用以及大大地扩展其对科学和社会的价值的理想还没有实现。要发挥散落在世界各地的物种信息的优势，就应当开展这项工作。事实上，保护生物学家、资源管理者及其他自然资源利用者曾多次要求获得这些信息。

本项任务中，应优先考虑以下几个方面的工作：

1. 以世界自然历史标本存储机构收藏的标本为基础，建立物种信息的系统学、生物地理学和生态学数据库。

2. 综合系统学标本存储机构的数据和地理信息系统(GIS)数据库的信息，为监测过去和现在全球变化对物种的分布和灭绝造成的影响提供条件。

3. 建立各数据库间的联系，使各种分类单元及其分布区所有有用的信息可有效地被查询。

4. 建立和采用一套信息系统，便于国际用户的使用。

5. 编撰所有系统学数据库必需的、包括分类名称、地理分布及其他信息内容的数据字典。

6. 出版手册、图解、电子动植物志、专著等数据产品。

7. 通过持续提供软件和硬件的支持，以建立维持和更新数据库及信息网络的机制。

有效的系统学信息系统的作用有：

1. 使政策制定者作出更全面的资源永续利用决策。
2. 更详细地记载物种的灭绝及其分布的变化。
3. 系统学及相关信息的数据库使用效率越高，就越能更经济地管理生物资源。
4. 方便系统学知识的获得，便于解决问题。
5. 在生物学及其他领域特别是生物技术的数据之间建立新的比较和联系的方法。
6. 加强国际交流与合作，减少科研工作的重复。

迎接挑战：基础设施与人才资源

被里约热内卢召开的联合国环境与发展大会以及生物多样性公约采纳的全球行动计划(21世纪议程)呼吁各参会国制定国家策略，去编目和了解本国的生物多样性，并制定今后的保护计划。在编目和了解世界生物多样性的过程中，国际社会将面临两个重大的挑战。第一，需要极力扩大和改善系统学研究的基础设施建设，特别是修盖生物标本收藏馆；第二，强化专业系统学人员的培训和录用，改变“分类学障碍”的说法。

系统学为生物多样性研究奠定了基础

“许多建议(加强基础和应用保护研究)认为，现有分类学知识能够承担这些工作。然而，完成这项工作所需要的训练有素的分类学队伍并不存在。生物多样性的描述、编目、分类、监察和管理知识必须经过培训。有了以上这些基础才能去研究和保护生物多样性。”

国家研究委员会，保护生物多样性，1992，第71页

系统学界通过“2000年系统学议程”提出系统学行动计划的建议。上面提到的三项任务的成功，对履行这个行动计划是必要的。这项计划有以下几个主要点：

1. 所有的国家为了解、保存和利用他们的生物多样性就要建立和加强具标本收藏馆的系统学研究中心；

2. 对系统学家和其支撑系统工作人员进行教育和培训的有关单位进行资助；

3. 在基础和应用系统学范围内要扩大研究工作的队伍；

4. 全世界系统学研究单位之间在研究和教育方面开展国际合作和交流；

5. 在上述这些单位之间，并扩大到社会上建立起电子交流的联系；

6. 支持基于分类单元和全世界范围的比较系统学研究。

建立和加强系统学研究中心及标本收藏

在一些国家中没有合适的基础设施来存放世界上大多数国家的生物多样性标本，是得不到生物多样性综合知识的一个原因。同时，发达国家的系统学研究中心也不是以支持它们为执行上述研究任务所承担的应负的责任。物种丰富国家未来的繁荣将有赖于发展管理他们自己的生物资源的能力，其中包括要具备为提供对有效地作出决策所需知识的科学基础设施。世界各国为克服科学能力的不足，就必须通过建造新的或加强现有标本收藏馆的基础设施，例如博物馆、标本馆和为微生物和遗传资源保存所需要的贮藏所。

全世界自然历史标本收藏馆拥有20亿号标本(Duckworth 等，1993)。我们虽然已掌握有如此巨大的生物学遗产，但是通过已保存的标本，只鉴定了地球物种多样性的一小部分。植物、动物和其它生物的系统学标本收藏馆仅仅是我们生物区系的永久性记录。起源于这些标本收藏馆的特殊的库和数据库是我们对地球自然历史的书面的记录(见专栏12)。已保存标本的系统学收藏馆可在自然历史博物馆、标本馆和大学的标本馆和农业部、自然资源或生物考察等政府有关部门的标本馆内见到。这些标本收藏馆也可以活生物的形式保存在动物园、水族馆、昆虫馆、鸟类馆、植物园内，或者如种质、冷冻的组织和微生物的模式标本保存在特殊的贮藏器内。

有保存标本的标本收藏(馆)的系统学研究中心馆是各国乃至全世界有关生物多样性知识的贮藏场所。尽管各国标本收藏的基础设施不同，但它们基本上具有两种主要功能。第一种是国家研究中心，通过标本馆、图书馆和数据库，记载本国的生物多样性。第二种设施也很有存在的必要，这类中心的功能是支持重点放在分类单元，且有国际意义的系统学研究或标本收藏。有些中心只有上述一种功能，而另外一些中心则具有足够的基础设施和科研力量，两种功能兼备。许多中心可以在现有机构和标本收藏馆的基础上建立，而有些则必须重新建立。不论是那种情况，科学设施的核心都应当包括有分类单元知识的专业系统学家。

专栏12

系统学标本收藏馆的重要性

"科学标本收藏馆是社会在了解自然界过程中的一项持续性投资……随着生境的消失、物种的灭绝和有重要地质学及古生物学价值的遗址的破坏，这些标本收藏馆中的标本已成为一种不可再生资源。"

—保存自然科学标本收藏馆：环境遗产记事，第6页

1. 标本收藏馆是人类自然遗产的永久性记录，它包含着许多科学领域的研究所不可缺少的材料，例如保存生物多样性和监测全球变化的研究；

2. 标本收藏馆能满足应用生物学的需求，例如卫生科学(寄生虫学、流行病学、诊断学)、农业、资源管理、生物技术等；

3. 标本收藏馆积极支持公共教育和正规教育计划；

4. 通过标本收藏馆的展览提高自然保护及生物多样性保护的公众意识。

国家系统学研究中心。国家研究中心应当建有收藏各地方或各地区动植物区系的标本收藏馆，使国内生物多样性研究和管理能够采用准确、最新的系统学数据。世界各地的标本收藏馆要求配有专业人员，以便提供当地生物区系鉴定准确的标本；这些标本许多都是当地的特有种。此外，全国生物学调查所得数据的核实也要依靠这些系统学标本收藏馆中的佐证标本。只有研究这些标本，研究人员才能断定同一物种在不同地区被发现。

国家生物资源研究中心的迅速发展以及世界各国制定的各种计划，如墨西哥的全国生物多样性交流委员会(CONABIO)、哥斯达黎加的国家生物多样性研究所(INBio)、美国的全国生物多样性调查等，都为此类项目在国内的顺利开展奠定了基础。这些计划作用重大，是研究、管理和保护各国生物多样性的基础，也是争取国内支持这些行动的依据。

国际系统学研究中心。2000年系统学议程行动计划的内容之一就是建立新的或完善现有的系统

学研究中心，为实现记录世界生物多样性这个目标作出贡献。各个国家研究中心重复各种生物类群的标本收藏馆或分类研究既不现实，也没有必要。相反，了解世界上某一分类单元情况的系统学家则必须得到收藏各国标本的研究中心的资助，以便在世界各地开展研究工作，因为物种往往经常是跨国境分布的。系统学研究最终以分类单元来定向的，所以就必须全面地在世界范围内收藏各种生物类群的标本。由于地区性标本收藏一般不够全面，很难广泛地满足比较系统学研究的要求，各国研究机构有必要充分地参与。由此，国际研究机构间的合作和数据库联网不可缺少。物种丰富国家研究机构间的伙伴关系也应当建立起来，以便保持系统学标本收藏馆的稳定和发展，建立起交流标本、生物多样性信息、培训水平和科学知识的网络。任何一个标本收藏馆都不可能有研究各种生物类群的系统学家，可见各类群世界级专家的人员交流也势在必行。系统学界如果想多快好省地编目和划分世界的物种多样性，建立起这种联系乃关键所在。

在各有关国家建立系统学研究中心，就要求开展一些重大的国际合作项目。除个别情况外，世界上物种丰富国家基本上没有标本收藏馆，即使有也得不到足够的支持。这些标本收藏馆大部分都缺乏专业系统学家和受过培训的辅助人员，可用于调查和编目的经费也普遍不足，更不用说资助系统学家到世界各地的标本收藏馆去进行必要的比较研究。物种丰富的国家如果要建立一套系统学知识体系，为生物多样性的保护和永续利用提供服务，就必须解决这些问题(见专栏13)。

在发达国家，系统学标本收藏馆的数量和侧重各不相同。以美国为例，主要的系统学研究中心目前有50多个，收藏从细菌到鲸几乎包括所有生命类型的标本。大型标本收藏中心都是一些重要的自然历史博物馆及植物园，州立机构、联邦机构、许多大学也有。此外，侧重收藏地区标本的研究机构也很多，这些标本收藏馆具有重要的历史和科学价值。

与系统学标本收藏馆有关的数据中心、图书馆和档案馆同样是系统生物学研究的重要场所。专业图书馆并不仅限于收藏书刊，也同样收集卡片索引、目录、原稿、插图、照片、缩微胶片记录、制图法资料、文献档案及不同类型的电子媒介物等。近几年来，科学信息的大量产生也相应使大多数研究机构的收集能力有所提高。如果生物多样性的全球调查和编目工作获得成功，即使是最大的研究中心，其贮存和管理信息的能力也会逐渐显得不足。只有加强基础设施建设，提高系统学数据库的储存、查询和利用的水平，才能使世界科学家共享信息，避免研究重复和面向社会。

专栏13

利用收藏标本解决问题

1. 当公共健康官员担心鱼类体内的水银含量时，可通过博物馆的标本样品了解水银污染的历史模式。

2. 60年代许多鸟类的种群数量开始下降，经研究博物馆收藏的鸟蛋，发现蛋壳日趋变薄，终于使研究人员联想到由于环境中存在有 DDT。

3. 导致爱滋病及其他疾病的病毒突变频繁，若与前曾描述过的、贮藏在模式菌种或冷冻标本中的菌株进行比较，便可帮助公共健康官员了解疾病的传播途径。

改善系统学基础设施的建议：

1. 建立和完善世界各国的国家系统学研究中心。这些中心应当盖有收藏记录国内生物财富的地方及地区参考标本收藏馆，具有供研究和储存信息之用的设施，并具备培训和教育能力；

2. 建立或完善国际系统学研究中心，以利于获得系统学知识和建设有分类单元研究重点的世界性标本收藏馆；

3. 通过建立国际网络，改善世界上标本的照管和保藏，实现知识、信息及资源的共享。

4. 加强和扩大现有系统学研究中心的储存能力及研究力量，包括制定相应计划，将所有已登记的生物多样性录入数据库。

5. 加强和扩大系统学研究中心间国际合作项目进行。

教育、培训及人才资源开发

在占世界80%的陆生生物多样性的国家中，其科学家数量仅占6%，这是获得科学知识、了解和有效利用生物多样性的一个严重制约。除此之外，另一个制约因素是全世界缺少许多生物类群的分类专家，包括一些最富有多样性和最有经济价值的类群。由此看来，实现2000年系统学议程提出的三项研究任务所面临的最大挑战，或许就是在全世界招募、教育、培训和聘用足够的系统学家和技术人员。美国国家科学理事会的全球生物多样性特别工作组(1989)、英国上议院的科学和技术选择委员会(1991)等众多组织和特别工作组的结论认为，缺少生物的鉴定、记录和分类方面的系统生物学家，严重妨碍了全球生物多样性减少之问题的有效解决。

训练有素的系统学家数量减少的原因很多。美国国家科学基金会近期指导进行的一项大学调查发现，在调查的有博士学位授予权的院校中，系统生物学家仅有940名，25%的所有调查院校只设置了副手职位(Higher Education Survey, 1990)。更可悲的是，收到的回复中只有18%的院校表明如果今后有新的教员职位，它们才会聘用系统学家。在美国和英国，系统学家的研究工作和(或)教学职位由于受到其他生物学领域的竞争，在几十年前已紧缩。因此，大部分研究机构在调查时提出它们愿扩大的是分子生物学而不是系统学项目，因为前者“具有更大的资助机会”。

出于上述原因，新培养的系统学家的数量比二十年前减少了很多。许多生物类群的分类专家越来越少，如藻类、细菌、真菌、低等无脊椎动物、昆虫及其近缘种等。一些对经济和生态系统最为重要的生物类群更需要专业培训的专家去作研究。以线虫为例，美国有15个机构所属的标本馆和12名专职线虫系统学家，但带学生的线虫系统学家只有两位。由于线虫对农业有举足轻重的作用，缺乏受过培训的专家必将会带来严重的经济后果。

改变“分类学障碍”的说法

“……过去几年中下降最严重的是分类学和系统学领域。受过培训的分类学家急缺，热带国家几乎没有。许多重要生物类群方面的专家也显不足，发达国家甚至也是如此。另外，可参考的标本收藏馆也不够，而且地区分布不平衡，大都远离被研究生物的原产地。目前分类学的效率低下，重复研究较多。有鉴于此，在开展任何生物多样性研究之前，必须首先强调改变‘分类学障碍’的说法。”

—DIVERSITAS: IUBS-IUMS-SCOPE-UNESCO

生物多样性项目

尽管人们对许多类群的系统学知识不足或严重不足有一致的看法，但尚欠缺能说明该学科有这种趋势的证据。统计世界现有的各生物类群的系统学专业人才十分必要，只有这样，才能更准确、更有效地重点实施培训和吸收人才计划。

系统生物学人才资源的匮乏在发展中国家更为突出。受过培训的科学家的人数、研究经费、标本馆职位、部门支持等方面的不足，都意味着很少有系统科学家投身这一学科行列。要扭转这种局势，迫切需要强化发展中国家地方自然历史研究机构的作用，建立与发达国家研究机构间的联系(National Research Council)。由于在很多发展中国家环境恶化和生物多样性丧失的趋势不断加快，培训“准分类学家”，让他们与专业系统学家一道工作也不失为解决人才资源问题的一种办法，但这并不意谓着可以取代建立有效的科学基础设施和强大的科学家队伍的计划。

在发达国家，不少像博物馆和植物园之类的研究机构开展了与附近大学联合培养系统生物学本科生和研究生的项目。这些项目一般说来都有很大的国际性成分。尽管取得的成绩不少，但就物种多样性的危机程度和描述世界生物资源在科学和经济上的紧迫性来看，系统学还必须得到更多的经费支持，争取在研究生和大学生中培养出更多的科学家。部分证据表明，大学生对物种多样性很感兴趣，他们为现代系统学研究的重要性和在智力上富有的挑战性所吸引。但由于研究生经费不足和博士生毕业后就业前景渺茫，他们只好打消投身系统学研究的念头。

加强人才资源的利用的建议：

1. 从事发展中国家持续发展和环境保护推动项目的国际机构应当为基础系统学研究以及专业、准专业系统学家的培训和聘任提供经费支持；

2. 各国应当建立全国生物多样性监测机构、系统学研究中心及标本收藏馆及国家生物普查机构，并为上述活动的开展配置专业的系统学家；

3. 各大学和研究所凡设有生物多样性培训和教育项目者，应当配备专业系统学教师。

4. 发达国家的政府机构和自然历史研究机构应当向物种丰富国家的学生及其他科研人员的培训提供更大的援助；

5. 资源管理机构（如林业、渔业、野生动物等）的工作人员中应当包括专业系统学家。

2000年系统学议程完善了其他生物多样性计划

国际社会中许多机构和组织已经认识到描述和了解地球物种多样性的紧迫性。这些机构和组织包括：

·联合国环境与发展大会（UNCED）及其二十一世纪议程

·联合国环境规划署（UNEP）

·联合国开发署（UNDP）

·Diversitas，一个包括以下机构和组织的联合体：

国际生物科学联盟（IUBS）

环境问题科学委员会（SCOPE）

联合国教科文组织（UNESCO）

国际微生物学联盟（IUMS）

·二十一世纪的微生物多样性，国际生物科学联盟（IUBS）和国际微生物学联盟（IUMS）的一项行动计划

非政府组织也发出同样的呼吁，特别是：

·世界资源研究所（WRI）

·世界自然基金会（WWF）

·国际自然保护联盟/国际自然与自然资源保护联盟（IUCN）

·大自然保护协会（TNC）

2000年系统学议程从分类单元着手提出了全球生物多样性的一个前景展望，给保护和管理这项必不可少资源的科学框架增添了新的内容。对所有旨在了解、保护和利用生物多样性的计划，如生态学界提出的持续性生物圈计划来说，2000年系统学议程起到了锦上添花的作用。

对2000年系统学议程的投资

2000年系统学议程之计划设想宏伟，需要众多方面的大力支持才能得以实现。任何政府或任何国际机构都不可能去单独资助这项计划的运作，它需要世界上众多组织和政府的共同支持。2000年系统学议程的任务通过一项长远的25年计划实现，每年大约需要投入经费30亿美元。

经费预算系根据各项工作的开支制订，其中包括人才开发、建立计算机网络、支持标本收藏建设（包括模式菌种及种质贮藏所）、研究项目、宣传研究成果等方面的开支。按照目前的资助水平，完成2000年系统学议程的任务至少需要150年。鉴于生物多样性丧失的紧迫性，目前每年的研究和基础建设资助（大约5亿美元）必须增加将近六倍才能在25年内完成这些任务。

从生物多样性显示出的科学、经济和美学价值中不难看出这项投资的产出。发现、描述、了解和利用其他物种取得的持续进展已经使人类受益匪浅。人类今天的进一步努力会带来明天的更大效益，同时也将有助于今后世世代代生命多样性的保存。

参考文献

Brooks, D. R. and B. McClennan. 1991. Phylogeny, ecology, and behavior: a research program in comparative biology. U. of Chicago Press, Chicago, Illinois.

Bull, A. T., M. Goodfellow, J. H. Slater. 1992. Biodiversity as a source of innovation in biotechnology. Annual Review of Microbiology. 46: 219—252.

di Castri, F., J. R. Vernhes, and T. Younes. 1992. Inventorying and monitoring biodiversity. Biology International, Special issue. 27, 1—28.

Duckworth, W. D., H. H. Genoways, and C. L. Rose, 1993. Preserving natural science collections: chronicle of our environmental heritage. National Institute for the Conservation of Cultural Property, Washington, D. C.

Ehrlich, P. R. and A. H. Ehrlich. 1992. The value of biodiversity. Ambio. 21: 219—226.

Ehrlich, P. R. and E. O. Wilson. 1991. Biodiversity studies: science and policy. Science. 253: 758—762.

Gaston, K. G. and R. M. May. 1992. Taxonomy of taxonomists. Nature. 365: 281—282.

Gore, A. 1992. Earth in the balance. Houghton Mifflin, New York.

Groombridge, B. (ed.). 1992. Global biodiversity: status of the Earth's living resources. World Conservation Monitoring Centre. Chapman and Hall, London.

Hawksworth, D. L. and J. M. Ritchie. 1993. Biodiversity and biosystematic priorities: microorganisms and invertebrates. International Mycological Institute, CAB International, Wallingford, United Kingdom.

Higher Education Surveys. 1990. Systematic biology training and personnel. Survey No. 10., Washington, D. C.

Holden, C. 1989. Entomologists wane as insects wax. Science. 246:754—756.

Iltis, H. H. 1989. Serendipity in the exploration of biodiversity: what good are weedy tomatoes? In Biodiversity (E. O. Wilson, ed.), National Academy Press, Washington, D. C., 98—105.

International Union for the Conservation of Nature (IUCN), World Resource Institute (WRI), Conservation International (CI), World Wildlife/U. S (WWF/US), and World Bank. 1990. Conserving the world's biological diversity.

Lubchenco, J., A. M. Olson, L. B. Brubaker, S. R. Carpenter, M. M. Holland, S. P. Hubbell, S. A. Levin, J. A. MacMahon, P. A. Matson, J. M. Melillo, H. A. Mooney, C. H. Peterson, H. R. Pulliam, L. A. Real, P. Regal, and P. G. Risser. 1991. The sustainable biosphere initiative: an ecological research agenda. Ecology. 72: 371—412.

May, R. M. 1992. How many species inhabit the Earth? Scientific American. 267(4): 42—48.

Miller, E. H. (ed). 1985. Museum collections: their role and future in biological research. Vancouver, B. C.: Brit. Columbia Prov. Museum Occas. Papers Series, No. 25.

Myers, N. 1988. Threatened biotas: "Hot spots" in tropical forests. Environmentalist. 8: 187—208.

Myers, N. 1990. The biodiversity challenge: Expanded hotspot analysis. Environmentalist. 10: 243—256.

Myers, N. 1991. Tropical forests: present status and future outlook. Climate Change. 19: 3—32.

Nash, S. 1989. The plight of systematists: are they an endangered species? Scientist Oct. 16: 7.

National Research Council. 1992. Conserving biodiversity: a research agenda for development agencies. National Academy Press, Washington, D. C.

National Research Council. 1993. A biological survey for the nation. National Academy Press, Washington, D. C.

National Science Board. 1989. Loss of biological diversity: a global crisis requiring international solutions. National Science Foundation, Washington, D. C.

Norse, E. A. (ed.). 1993. Global marine biological diversity. Island Press, Washington, D. C.

Office of Technology Assessment (OTA). 1987. Technologies to maintain biological diversity. Washington, D. C.

Peters, R. L., and T. E. Lovejoy (eds.). 1992. Global warming and biological diversity. Yale University Press, New Haven, Connetiicut.

Rasoanaivo, P. 1990. Rain forests of Madagascar: sources of industrial and medicinal plants. Ambio. 19: 421—424.

Raven, P. H. 1993. A plea to the citizens of the world: live as if Earth matters. Diversity. 9(3): 49—51.

Raven, P. H. and E. O. Wilson. 1992. A fifty-year plan for biodiversity studies. Science. 258: 1099—1100.

Reid, W. V., and K. R. Miller. 1989. Keeping options alive: the scientific basis for conserving biodiversity. World Resources Institute, Washington, D. C.

Resolutions from the International Symposium and First World Congress on the Preservation and Conservation of Natural History Collections. 1992. Madrid, Spain, 1—57; EBCOMP, S. A., Bergantin, 1—28042. Madrid. Endorsed by United Nations Education, Scientific and Cultural Organization.

Sitarz, D. (ed.). 1993. AGENDA 21: the Earth Summit strategy to save our planet. Earthpress, Boulder, Colorado.

Stierle, A., G. Strobel, and D. Stierle. 1993. Taxol and taxane production by taxomyces andreanae, an endophytic fungus of Pacific Yew. Science. 260: 214—216.

Task Force on Global Biodiversity. 1989. Loss of biological diversity: a global crisis requiring international solutions. NSF—89—171, Washington, D. C.

U. K. House of Lords. 1991. Systematic Biology Research. Select Committee on Science and Technology, 1st Report. London: HMSO, HL Paper 22—I.

U. K. Natural Environment Research Council. 1992. Evolution and biodiversity -the new taxonomy. London.

United Nations Conference on Environment and Development (UNCED). 1992. United Nations Convention on Biological Diversity.

Wheeler, Q. D. 1990. Insect diversity and cladistic constraints. Annals of the Entomological Society of America. 83: 1031—1047.

Wilson, E. O. 1985. The biological diversity crisis: a challenge to science. Issues Sci. Technol. (Fall): 20—29.

Wilson, E. O. (ed.). 1988. Biodiversity. National Academy Press, Washington, D. C.

Wilson, E. O. 1992. The diversity of life. Harvard University Press, Cambridge, Ma.

Wilson, E. O. 1992. Biodiversity: challenge, science, opportunity. American Zoologist. 32: 1—7.

World Resources Institute. 1993. Biodiversity prospecting: using genetic resources for sustainable development. World Resources Institute, Washington, D. C.

World Resources Institute (WRI), International Union for the conservation of Nature (IUCN), and United Nations Environment Program (UNEP). 1992. Global Biodiversity Strategy. Guidelines for action to save, study, and use the earth's biotic wealth sustainably and equitably. WRI, Washington, D. C.

词　汇

生物多样性(Biodiversity)：生物的多样性和变异性，包括物种的数量、独特分支(distinct clades)的数量、种内遗传性变异及"功能的"多样性(生物的功能、生物与其他生物及其环境间相互作用的多得不计其数的方式)。

生物地理学(Biogeography)：对植物、动物和微生物地理分布的研究。

生物资源(Biological Resource，Bioresources)：对人类有实在或潜在价值的某种生物，或者源于某种生物的产品。

生物圈(Biosphere)：有生命存在的地球部分。

生物技术(Biotechnology)：利用生物或源于生物的物质，去制造或改变某种产品，改良植物或动物，或者为特定目的去开发微生物的任何技术。

分支(Clade)：某一分类单元，系统发育中的一个分枝或一个谱系。

分类(Classification)：正式地、科学地把物种排列到某种等级系统，并给物种或物种类群制订科学名称。

比较生物学(Comparative biology)：一个以上物种之间的比较研究。

特有(Endemic)：某种生境或地理区域特有的某一物种。

绝灭(Extinction)：某一物种最后一个个体的死亡。

编目(Inventory)：把某一指定区域的所有植物、动物及微生物种的名称或描述编列成表。

专著(Monograph)：汇集世界上关于某一分类单元的所有已知资料，包括新的及过去已知的物种，按照已知的系统发育组织编写的一本全面的论著。

系统发育多样性(Phylogenetic diversity)：某一类群所代表的独特的谱系的数量或分支的数量。

系统发育(Phylogeny)：物种间进化史的模式及物种间的共同祖先。

物种(Species)：生物的种类；分类学和系统发育分析的基本单位。

物种多样性(Species diversity)：原义是指即地球上物种的数量和种类；广义上亦包括物种的数量及其相互关系(如系统发育多样性)。

调查(Survey)：按照一定的方法考察某一选定地区，以便发现大型或微型的动植物区系内的所有物种。

系统学(Systematics)：现在生活的和化石生物(物种)种类的比较研究，包括它们的描述、分布及其共存的系统发育关系。

分类单元(Taxon)：一个物种或一组相关的物种。

分类学(Taxonomy)：把物种科学地划分到一种等级系统，以此反映对其系统发育的了解情况。

(郭寅峰译　钱迎倩和周红章校)

ISBN 7-03-005669-8

9 787030 056696 >

ISBN 7-03-005669-8
定价:41.00 元